

## BIROn - Birkbeck Institutional Research Online

Nocera, Andrea (2017) Estimation and inference in mixed fixed and random coefficient panel data models. Working Paper. Birkbeck, University of London, London, UK.

Downloaded from: <https://eprints.bbk.ac.uk/id/eprint/26864/>

*Usage Guidelines:*

Please refer to usage guidelines at <https://eprints.bbk.ac.uk/policies.html>  
contact [lib-eprints@bbk.ac.uk](mailto:lib-eprints@bbk.ac.uk).

or alternatively

ISSN 1745-8587



Department of Economics, Mathematics and Statistics

BWPEF 1703

# **Estimation and Inference in Mixed Fixed and Random Coefficient Panel Data Models**

Andrea Nocera  
*Birkbeck, University of London*

June 2017

# Estimation and Inference in Mixed Fixed and Random Coefficient Panel Data Models

Andrea Nocera  
Birkbeck, University of London\*

June, 2017

## Abstract

In this paper, we propose to implement the EM algorithm to compute restricted maximum likelihood estimates of both the average effects and the unit-specific coefficients as well as of the variance components in a wide class of heterogeneous panel data models. Compared to existing methods, our approach leads to unbiased and more efficient estimation of the variance components of the model without running into the problem of negative definite covariance matrices typically encountered in random coefficient models. This in turn leads to more accurate estimated standard errors and hypothesis tests. Monte Carlo simulations reveal that the proposed estimator has relatively good finite sample properties. In evaluating the merits of our method, we also provide an overview of the sampling and Bayesian methods commonly used to estimate heterogeneous panel data. A novel approach to investigate heterogeneity of the sensitivity of sovereign spreads to government debt is presented.

**JEL Classification:** C13, C23, C63, F34, G15, H63.

**Keywords:** EM algorithm, restricted maximum likelihood, correlated random coefficient models, heterogeneous panels, debt intolerance, sovereign credit spreads.

---

\*I am very grateful to Ron Smith and Zacharias Psaradakis for their careful reading, comments and advice. I would also like to thank Hashem Pesaran, Ivan Petrella, and the participants at the “New Trends and Developments in Econometrics” 2016 conference organized by the Banco de Portugal, at the IAAE 2016 annual conference, at Bristol Econometric Study Group 2016, CFE-CMStatistics 2016, SAEe 2016, and RES PhD Meetings 2017, and the seminar participants at Ca’ Foscari University of Venice, Carlos III University of Madrid (Department of Statistics), University of Kent, and University of Southern California USC Dornsife INET. I am also thankful to the seminar participants at the Deutsche Bundesbank, and at Birkbeck in 2015 for their comments on a preliminary version of this paper. All remaining errors are mine.  
E-mail address: a.nocera@mail.bbk.ac.uk

# 1 Introduction

This paper considers the problem of statistical inference in random coefficient panel data models, when both  $N$  (the number of units) and  $T$  (the number of time periods) are quite large. In the presence of heterogeneity, the parameters of interest may be the unit-specific coefficients, their expected values, and their variances over the units. Two main estimators for the expected value of the random coefficients are used in the literature. Pesaran and Smith (1995) suggest estimating  $N$  time series separately to then obtain an estimate of the expected value of the unit-specific coefficients by averaging the OLS estimates for each unit. They call this procedure Mean Group estimation. Alternatively, under the assumption that the coefficients are random draws from a common distribution, one can apply Swamy (1970) GLS estimation, which yields a weighted average of the individual OLS estimates.<sup>1</sup> However, as in the error-component model, the Swamy estimator of the random coefficient covariance matrix is not necessarily nonnegative definite. Our aim is to investigate the consequences of this drawback in finite samples, in particular when testing hypotheses. At the same time, we propose a solution to the above mentioned problem by applying the EM algorithm. In particular, following the seminal papers of Dempster et al. (1977), and Patterson and Thompson (1971), we propose to estimate heterogeneous panels by applying the EM algorithm to obtain tractable closed form solutions of restricted maximum likelihood (REML) estimates of both fixed and random components of the regression coefficients as well as the variance parameters. The proposed estimation procedure is quite general, as we consider a broad framework which incorporates various panel data models as special case. Our approach yields an estimator of the average effects which is asymptotically related to both the GLS and the Mean Group estimator, and which performs relatively well in finite sample as shown in our limited Monte Carlo analysis. We also review some of the existing sampling and Bayesian methods commonly used to estimate heterogeneous panel data, to highlight similarities and differences with the EM-REML approach.

Both the EM and the REML are commonly used tools to estimate linear mixed models but have been neglected by the literature on panel data with random coefficients.<sup>2</sup> The EM

---

<sup>1</sup>Swamy focuses on estimating the average effects while the random effects are treated as nuisance parameters and conditioned out of the problem. However, the estimation of the random components of the model becomes crucial if the researcher wishes to predict future values of the dependent variable for a given unit or to describe the past behavior of a particular individual. Joint estimation of the individual parameters and their mean has been proposed by Lee and Griffiths (1979). Joint estimation in a Bayesian setting has been suggested by Lindley and Smith (1972), and has been further studied by Smith (1973), Maddala et al. (1997) and Hsiao, Pesaran and Tahmiscioglu (1999). A good survey of the literature is provided by Hsiao and Pesaran (2008), and Smith and Fuertes (2016).

<sup>2</sup>For discussions on EM and REML estimation of linear mixed models, see Hariville (1977), Searle and Quaas (1978), Laird and Ware (1982), Pawitan (2001), and McLachlan and Krishnan (2008), among others.

algorithm has also recently gained attention in the finance literature. Harvey and Liu (2016) suggest a similar approach to ours to evaluate investment fund managers. The authors focus on estimating the fund-specific random effects population (“alphas”) while the other coefficients of the model (“betas”) are assumed to be fixed. Instead, we consider a different framework where both the intercept and slope parameters are a function of a set of explanatory variables and are randomly drawn from a common distribution. We derive an expression for the likelihood of the model accordingly. More importantly, differently from Harvey and Liu, our goal is to illustrate the advantages of the EM-REML approach in estimating a general class of heterogeneous panel data models, in relation to the existing methods.

First, estimating heterogeneous panels by EM-REML yields unbiased and more efficient estimation of the variance components. This is important as the unbiased estimator of the variance-covariance matrix of the random coefficients proposed by Swamy (1970) is often negative definite. In such cases, the author suggests eliminating a term to obtain a non-negative definite matrix. This alternative estimator is consistent when  $T$  tends to infinity but it is severely biased in small samples. As shown in the Monte Carlo analysis, this in turn leads to biased estimated standard errors and may affect the power performances of the GLS estimator. Compared to Swamy estimator, the EM-REML method leads to remarkable reduction of the bias and root mean square errors of the estimates of the random coefficient variances. As a results, the estimated standard errors have lower bias, leading to more accurate hypothesis tests. A valid estimator of the random coefficient covariance matrix is also important to correctly detect the degree of coefficient heterogeneity. As noted by Trapani and Urga (2009), the latter plays a crucial role on the forecasting performance of various panel estimators, while other features of the data have a very limited impact. Therefore, our estimator of the covariance matrix may be considered by applied researchers to choose the appropriate estimator for forecast purposes.

Lee and Griffiths (1979) derive a recursive system of equations as a solution to the maximization of the likelihood function of the data which incorporates the prior likelihood of the random coefficients. However, we demonstrate that their estimate of the random coefficients’ variance-covariance matrix does not satisfy the law of total variance. This is not the case when using the EM algorithm. Differently from Lee and Griffiths, we consider the joint likelihood of the observed data and the random coefficients as an incomplete data problem (in a sense which will be more clear later on). We show that maximizing the expected value of the joint likelihood function with respect to the conditional distribution of the random effects given the observed data is necessary for the law of total variance to hold.

Another interesting feature of the EM (compared to the papers mentioned in the above paragraph) is that it allows us to make inference on the random effects’ population. Indeed, in general, it gives a probability distribution over the missing data.

The random effects are estimated by the mean of their posterior distribution, under the

assumption that the regressors are strictly exogenous. Substituting the unknown variance components by their estimates yields the empirical best linear unbiased predictor. We also note that the EM-REML estimator of the average effects is related to the empirical Bayesian estimator described in Hsiao, Pesaran and Tahmiscioglu (1999). The EM-REML estimators of the variance components are analogous to the Bayes mode of their posterior distribution, derived in Lindley and Smith (1972). In view of the relatively good finite-sample performances, the EM approach should be regarded as a valid alternative to Bayesian estimation in those cases in which the researcher wishes to make inference on the random effects distribution while having little knowledge on what sensible priors might be. At the same time, a drawback of the Bayesian approach is that, when sample sizes is not too large (relative to the number of parameters being estimated), the prior choice will have a heavy weight on the posterior, which will consequently be far from being data dominated (Kass and Wasserman, 1996). To illustrate, Hsiao, Pesaran and Tahmiscioglu (1999), suggest using the Swamy covariance's estimator as a prior input for the random coefficient covariance matrix. However, they note that the latter affects the empirical and hierarchical Bayes estimates of the regression coefficients adversely, especially when the degree of coefficient heterogeneity decreases. Alternatively, when considering a diffuse prior, their Gibbs sampling algorithm breaks down completely in some experiments. Another merit of our method is to overcome this problem.

The proposed econometric methodology is used to study the determinants of the sensitivity of sovereign spreads with respect to government debt. While there is a large literature on the empirical determinants of sovereign yield spreads there is no work, to the best of our knowledge, which tries to explain and quantify the cross-sectional difference in the reaction of sovereign spreads to change in government debt.<sup>3</sup> First, we show that financial markets reactions to an increase in government debt are heterogeneous. We then model such reactions as function of macroeconomic fundamentals and a set of explanatory variables which reflect the history of government debt and economic crises of various forms. We find that country-specific macroeconomic indicators, commonly found to be significant determinants of sovereign credit risk, do not have any significant impact on the sensitivity of spreads to debt. On the other hand, history of repayment plays an important role. A 1% increase in the percentage of years in default or restructuring domestic debt is associated with around 0.35% increase in the additional risk premium in response to an increase in debt.

The paper is organized as follows. Section 2 describes the regression model and its main assumptions. In Section 3 an expression for the likelihood of the complete data, which includes both the observed and the missing data, is obtained. The restricted likelihood is also derived. Section 4 illustrates the use of EM algorithm and shows how to perform the two

---

<sup>3</sup>The effects of macroeconomic fundamentals on sovereign credit spreads are examined in Akitoby and Stratmann (2008), Bellas et al. (2010), Edwards (1984), Eichengreen and Mody (2000) and Hilscher and Nosbusch (2010), among others.

steps of the EM algorithm, called the E-step and the M-step. We compare the EM-REML approach with alternative methods in Section 5. The problem of inference in finite sample is addressed in Section 6. Results from Monte Carlo experiments are shown in Section 7. In Section 8, we employ the econometric model to study the determinants of the sensitivity of sovereign spreads. Finally, we conclude.

## 2 A Mixed Fixed and Random Coefficient Panel Data Model

We assume that the dependent variable,  $y_{it}$ , is generated according to the following linear panel model with unit-specific coefficients,

$$y_{it} = c_i + x'_{it}\beta_i + \varepsilon_{it}, \quad (1)$$

for  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , where  $x_{it}$  is a  $K \times 1$  vector of exogenous regressors. The model can be written in stacked form

$$y_i = Z_i\psi_i + \varepsilon_i, \quad (2)$$

where  $y_i$  is a  $T \times 1$  vector of dependent variables for unit  $i$ , and  $Z_i$  is a  $T \times K^*$  matrix of explanatory variables, including a vector of ones.<sup>4</sup> Following Hsiao et al. (1993), in order to provide a more general framework which incorporates various panel data models as special case, we partition  $Z_i$  and  $\psi_i$  as

$$Z_i = \begin{bmatrix} \bar{Z}_i & \underline{Z}_i \end{bmatrix}, \quad \psi_i = \begin{bmatrix} \psi_{1i} \\ \psi_{2i} \end{bmatrix},$$

where  $\bar{Z}_i$  is  $T \times k_1^*$  and  $\underline{Z}_i$  is  $T \times k_2^*$ , with  $K^* = k_1^* + k_2^*$ . The coefficients in  $\psi_{1i}$  are assumed to be constant over time but differ randomly across units. Individual-specific characteristics are the main source of heterogeneity in the parameters:

$$\psi_{1i} = \Gamma_1 f_{1i} + \gamma_i, \quad (3)$$

where  $\gamma_i$  is a  $k_1^* \times 1$  vector of random effects,  $\Gamma_1$  is a  $(k_1^* \times l_1)$  matrix of unknown fixed parameters, and  $f_{1i}$  is a  $l_1 \times 1$  vector of observed explanatory variables that do not vary over time (for instance, Smith and Fuertes (2016) suggest using the group means of the  $x_{it}$ 's). The first element of  $f_{1i}$  is one to allow for an intercept. The coefficients of  $\underline{Z}_i$  are non-stochastic and subject to

$$\psi_{2i} = \Gamma_2 f_{2i}, \quad (4)$$

---

<sup>4</sup>To make notation easier, we assume that  $T = T_i$ , for all  $i$ , although the results are also valid for an unbalanced panel.

where  $\Gamma_2$  is a  $(k_2^* \times l_2)$  matrix of unknown fixed parameters, and  $f_{2i}$  is a  $l_2 \times 1$  vectors of observed unit-specific characteristics. Equations (3) and (4) can be rewritten as

$$\begin{aligned}\psi_{1i} &= \left( f'_{1i} \otimes I_{k_1^*} \right) \bar{\Gamma}_1 + \gamma_i, \\ \psi_{2i} &= \left( f'_{2i} \otimes I_{k_2^*} \right) \bar{\Gamma}_2,\end{aligned}\tag{5}$$

where  $\bar{\Gamma}_j = \text{vec}(\Gamma_j)$ , which is a  $k_j^* l_j$ -dimensional vector and  $F_{ji} = \left( f'_{ji} \otimes I_{k_j^*} \right)$  is a  $k_j^* \times k_j^* l_j$  matrix, for  $j = 1, 2$ . Substituting (5) into (2) yields

$$y_i = W_i \bar{\Gamma} + \bar{Z}_i \gamma_i + \varepsilon_i,\tag{6}$$

for  $i = 1, \dots, N$ , where

$$\begin{aligned}W_i &= \begin{bmatrix} \bar{Z}_i F_{1i} & \bar{Z}_i F_{2i} \end{bmatrix}, & \bar{\Gamma} &= \begin{bmatrix} \bar{\Gamma}_1 \\ \bar{\Gamma}_2 \end{bmatrix}, \\ T \times \bar{K} & & \bar{K} \times 1 & \end{aligned}$$

with  $\bar{K} = (k_1^* l_1 + k_2^* l_2)$ . We assume that:

(i) The regression disturbances are independently distributed with zero means and variances that are constant over time but differ across units:

$$\varepsilon_{it} \sim IIN(0, \sigma_{\varepsilon_i}^2).\tag{7}$$

(ii)  $x_{it}$  and  $\varepsilon_{is}$  are independently distributed for all  $t$  and  $s$  (i.e.  $x_{it}$  are strictly exogenous). Both set of variables are independently distributed of  $\gamma_j$ , for all  $i$  and  $j$ .

(iii)  $f_{1i}$  and  $f_{2i}$  are independent of the  $\varepsilon_{jt}$ 's and  $\gamma_j$ , for all  $i$  and  $j$ .

(iv) The vector of unit-specific random effects is independently normally distributed as<sup>5</sup>

$$\gamma_i \sim IIN(0, \Delta), \quad \forall i.\tag{8}$$

**Special Cases.** Many panel data models can be derived as special cases of the model described in equation (6). Among others:

1. Models in which all the coefficients are stochastic and depend on individual-specific characteristics can be obtained from (6) by setting  $\bar{Z}_i = 0$ .
2. Swamy (1970) random coefficients model requires  $\bar{Z}_i = 0$ , and  $f_{1i} = 1$ , for all  $i = 1, \dots, N$ , while  $\bar{\Gamma} = \psi$  is a  $K^* \times 1$  vector of fixed coefficients.

---

<sup>5</sup>In a previous version of this paper we noted that in our setting one can easily allow for heteroskedasticity of unknown functional form, by letting  $\text{var}(\gamma_i | f_i)$  to be different from  $\text{var}(\gamma_j | f_j)$ , for  $i \neq j$ .



3. The correlated random effects (CRE) model proposed by Mundlak (1978) and Chamberlain (1982) can be obtained by setting  $\bar{Z}_i = \iota$  (where  $\iota$  is a vector of ones),  $f_{1i}$  contains  $\bar{x}_i$ , the average over time of the  $x_{it}$ 's;  $f_{2i} = 1$  for all  $i$ , which implies that  $\psi_{2i} = \psi_2$  is a vector of common coefficients..
4. Error-components models (as described in Baltagi (2005) and in Hsiao (2003)) which are a special case of the CRE model with  $f_{1i} = 1$  for all  $i$  and  $\Gamma_1 \equiv c \in \mathbb{R}$ .
5. Model with interaction terms (e.g. Friedrich (1982)):  $\bar{Z}_i = 0$  and for instance  $f_{2i} = 1$ , while  $\underline{Z}_i$  contains the interaction terms.
6. Common Model for all cross-sectional units:  $\bar{Z}_i = 0$ , and  $f_{2i} = 1$  for all  $i$ .<sup>6</sup>

### 3 Likelihood of the Complete Data

Define the full set of (fixed) parameters to be estimated as

$$\theta = (\bar{\Gamma}', \sigma_\varepsilon^2, \omega')' = (\theta_1', \omega')',$$

where  $\sigma_\varepsilon^2 = (\sigma_{\varepsilon 1}^2, \dots, \sigma_{\varepsilon N}^2)$  and  $\omega$  is a vector containing the non-zero elements of the covariance matrix  $\Delta$ . We consider the unobserved random effects,  $\gamma = (\gamma_1', \dots, \gamma_N')'$ , as the vector of missing data, and  $(y', \gamma')'$  as the complete data vector. Following the rules of probability, the log-likelihood of the complete data is given by

$$\log L(y, \gamma; \theta) = \log f(y \mid \gamma; \theta_1) + \log f(\gamma; \omega), \quad (9)$$

which is the sum of the conditional log-likelihood of the observed data and the log-likelihood of the missing data.<sup>7</sup> Using assumption (8), the joint log-likelihood of the vector of missing data can be written as

$$\log f(\gamma) = \sum_{i=1}^N \log f(\gamma_i) = \mu_1 + \frac{N}{2} \log |\Delta| - \frac{1}{2} \sum_{i=1}^N \gamma_i' \Delta^{-1} \gamma_i. \quad (10)$$

We now derive the likelihood of  $y = (y_1', \dots, y_N')'$  given  $\gamma$ . From (6) we can easily obtain the conditional expectation and variance of  $y_i$ , which are given by  $E(y_i \mid \gamma_i) = W_i \bar{\Gamma} + \bar{Z}_i \gamma_i$  and  $\text{var}(y_i \mid \gamma_i) = \text{var}(\varepsilon_i) = R_i = \sigma_{\varepsilon_i}^2 I_T$ , respectively. Under the assumption that both the regression error terms,  $\varepsilon_i$ , and the random effects,  $\gamma_i$ , are independent and normally

---

<sup>6</sup>Models 5 and 6 do not involve any random coefficient and do not require the use of the EM algorithm.

<sup>7</sup>To make notation easier, hereafter, we write  $f(\gamma)$  and  $f(y \mid \gamma)$  instead of  $f(\gamma; \omega)$  and  $f(y \mid Z, \gamma; \theta_1)$  respectively.

distributed, it follows that  $y_i$  is normally distributed and independent of  $y_j$ , for  $i \neq j$ . Therefore, the conditional log-likelihood of the observed data is given by

$$\log f(y | \gamma) = \sum_{i=1}^N \log f(y_i | \gamma_i) = \mu_2 - \frac{1}{2} \sum_{i=1}^N \log |R_i| - \frac{1}{2} \sum_{i=1}^N \varepsilon_i' R_i^{-1} \varepsilon_i, \quad (11)$$

where

$$\varepsilon_i = y_i - W_i \bar{\Gamma} - \bar{Z}_i \gamma_i. \quad (12)$$

Having found an explicit formulation for  $\log f(y | \gamma; \theta_1)$  and  $\log f(\gamma; \omega)$ , we can derive an expression for the log-likelihood of the complete data by substituting (11) and (10) into (9). At this point, we can make two important observations. First,  $\theta_1$  and  $\omega$  are not functionally related (in the sense of Hayashi (2000, Section 7.1)). This implies that  $\log f(\gamma; \omega)$  does not contain any information about  $\theta_1$  and similarly  $\log f(y | \gamma; \theta_1)$  does not contain any information about  $\omega$ . Second, as stated in Harville (1977), the maximum likelihood estimation takes no account of the loss in degrees of freedom that results from estimating the fixed coefficients, leading to a biased estimator of  $\sigma_\varepsilon^2$ . In the next subsection, we eliminate this problem by using the restricted maximum likelihood (REML) approach, described formally by Patterson and Thompson (1971).

### 3.1 Restricted Likelihood

Following Patterson and Thompson (1971), we can separate  $\log f(y_i | \gamma_i; \theta_1)$  in two parts:  $L_{1i}$  and  $L_{2i}$ . By maximizing the former, we can estimate  $\sigma_{\varepsilon_i}^2$ . An estimate of  $\bar{\Gamma}$  is obtained after maximizing  $L_{2i}$ . The two parts can be obtained by defining two matrices  $S_i$  and  $Q_i$  such that the likelihood of  $(y_i | \gamma_i)$  (for  $i = 1, \dots, N$ ) can be decomposed as the product of the likelihoods of  $S_i y_i$  and  $Q_i y_i$ , i.e.

$$\log f(y_i | \gamma_i; \theta_1) = L_{1i} + L_{2i}. \quad (13)$$

Such matrices must satisfy the following properties: (i) the rank of  $S_i$  is not greater than  $T - \underline{K}$ , while  $Q_i$  is a matrix of rank  $\underline{K}$ , (ii)  $L_{1i}$  and  $L_{2i}$  are statistically independent, i.e.  $\text{cov}(S_i y_i, Q_i y_i) = 0$ , (iii) the matrix  $S_i$  is chosen so that  $E(S_i y_i) = 0$ , i.e.  $S_i W_i = 0$ , and (iv) the matrix  $Q_i W_i$  has rank  $\underline{K}$ .<sup>8</sup>

**Finding an expression for  $L_{1i}$ .** Premultiplying both sides of (6) by  $S_i$ , we have  $E(S_i y_i | \gamma_i) = S_i \bar{Z}_i \gamma_i$ , since  $S_i W_i = 0$  and  $\text{var}(S_i y_i | \gamma_i) = S_i R_i S_i'$ . Therefore, the conditional log-likelihood of  $S_i y_i$  is given by

$$L_{1i} = \mu_3 - \frac{1}{2} \log |S_i R_i S_i'| - \frac{1}{2} (y_i - \bar{Z}_i \gamma_i)' S_i' (S_i R_i S_i')^{-1} S_i (y_i - \bar{Z}_i \gamma_i). \quad (14)$$

---

<sup>8</sup>  $\underline{K} = \text{rank}(W_i) \leq \bar{K} < T$ .

Searle (1978) showed that “it does not matter what matrix  $S_i$  of this specification we use; the differentiable part of the log-likelihood is the same for all  $S_i$ ’s”. In other words, the log-likelihood  $L_{1i}$  can be written without involving  $S_i$ .<sup>9</sup> Indeed, equation (14) can be rewritten as

$$L_{1i} = \mu_3 - \frac{1}{2} \log |R_i| - \frac{1}{2} \log |W_i' R_i^{-1} W_i| - \frac{1}{2} \bar{\varepsilon}_i' R_i^{-1} \bar{\varepsilon}_i, \quad (15)$$

where  $\bar{\varepsilon}_i = y_i - W_i \hat{\bar{\Gamma}} - \bar{Z}_i \gamma_i$ , and  $\hat{\bar{\Gamma}}$  denotes the generalized least squares (GLS) estimator of  $\bar{\Gamma}$ , which we describe in Subsection 4.4.

**Finding an expression for  $L_{2i}$ .** Following Patterson and Thompson (1971), we can set  $Q_i = W_i' R_i^{-1}$  since it satisfies  $\text{cov}(S_i y_i, Q_i y_i) = 0$ . After premultiplying both sides of (6) by  $Q_i$ , we have  $E(Q_i y_i | \gamma_i) = W_i' R_i^{-1} (W_i \bar{\Gamma} + Z_i \gamma_i)$  and  $\text{var}(Q_i y_i | \gamma_i) = W_i' R_i^{-1} W_i$ . The log-likelihood of  $Q_i y_i | \gamma_i$  is given by

$$L_{2i} = \mu_4 - \frac{1}{2} \log |W_i' R_i^{-1} W_i| - \frac{1}{2} \varepsilon_i' H_i \varepsilon_i, \quad (16)$$

where  $H_i = R_i^{-1} W_i (W_i' R_i^{-1} W_i)^{-1} W_i' R_i^{-1}$  and the  $\varepsilon_i$ ’s are the regression errors defined in (12).

## 4 EM-Algorithm

### 4.1 Generalities

Using equations (9), (10) and (11), the log-likelihood of the complete data can be rewritten as

$$\begin{aligned} \log L(y, \gamma; \theta) &= \sum_{i=1}^N \{\log L(y_i, \gamma_i; \theta)\} \\ &= \sum_{i=1}^N \{\log f(y_i | \gamma_i; \theta_1) + \log f(\gamma_i; \omega)\}. \end{aligned}$$

Lee and Griffiths (1979) obtain iterative estimates of  $\theta$  and  $\gamma$  by maximizing directly the latter. Instead, we argue in favour of using the EM algorithm to compute maximum likelihood estimates as this method has some added advantages. First, as established in Dempster et al. (1977), the EM algorithm assures that each iteration increases the likelihood. Second, as it will be shown in the next sections, contrary to Lee and Griffiths approach which delivers

---

<sup>9</sup>Detailed derivations of  $L_{1i}$  and  $L_{2i}$  are described in Appendix A.1.

$var\{E(\gamma_i | y_i)\}$  as an estimator of  $var(\gamma_i)$ , the unconditional variance of the  $\gamma_i$ , the EM algorithm yields an estimator of the latter satisfying the law of total variance. Finally, the EM allows us to make inference on the random effects' population.

Moreover, to obtain unbiased estimates of the variances of the time-varying disturbances, we consider the complete-data (restricted) log-likelihood:

$$\log L(y_i, \gamma_i; \theta) = L_{1i} + L_{2i} + \log f(\gamma_i; \omega_i), \quad (17)$$

for  $i = 1, \dots, N$ , where  $\log f(y_i | \gamma_i; \theta_1)$  has been decomposed as shown in equation (13).

On each iteration of the EM algorithm, there are two steps. The first step, called E-step, consists in finding the conditional expected value of the complete-data log-likelihood.

Let  $\theta^{(0)}$  be some initial value for  $\theta$ . On the  $b$ th iteration, for  $b = 1, 2, \dots$ , the E-step requires computing the conditional expectation of the  $\log L(y, \gamma; \theta)$  given  $y$ , using  $\theta^{(b-1)}$  for  $\theta$ , which is given by

$$\begin{aligned} Q &= Q(\theta; \theta^{(b-1)}) = E_{\theta^{(b-1)}} \{ \log L(y, \gamma; \theta) | y \} \\ &= \sum_{i=1}^N E_{\theta^{(b-1)}} \{ \log L(y_i, \gamma_i; \theta) | y_i \} = \sum_{i=1}^N Q_i, \end{aligned} \quad (18)$$

where

$$Q_i = Q_i(\theta; \theta^{(b-1)}) \equiv E_{\theta^{(b-1)}} \{ \log L(y_i, \gamma_i; \theta) | y_i \} = Q_{1i} + Q_{2i} + Q_{3i},$$

and

$$\begin{aligned} Q_{1i} &= E_{\theta^{(b-1)}} \{ L_{1i} | y_i \}, \\ Q_{2i} &= E_{\theta^{(b-1)}} \{ L_{2i} | y_i \}, \\ Q_{3i} &= E_{\theta^{(b-1)}} \{ \log f(\gamma_i; \omega) | y_i \}. \end{aligned} \quad (19)$$

In practice, we replace the missing variables, i.e. the random effects ( $\gamma_i$ ), by their conditional expectation given the observed data  $y_i$  and the current fit for  $\theta$ .

The second step (M-Step) consists of maximizing  $Q(\theta; \theta^{(b-1)})$  with respect to the parameters of interest,  $\theta$ . That is, we choose  $\theta^{(b)}$  such that  $Q(\theta^{(b)}; \theta^{(b-1)}) \geq Q(\theta; \theta^{(b-1)})$ . In other words, the M-step chooses  $\theta^{(b)}$  as

$$\theta^{(b)} = \arg \max_{\theta} Q(\theta; \theta^{(b-1)}).$$

Starting from suitable initial parameter values, the E- and M-steps are repeated until convergence, i.e. until the difference  $L(y; \theta^{(b)}) - L(y; \theta^{(b-1)})$  changes by an arbitrarily small amount, where  $L(y; \theta)$  denotes the likelihood of the observed data.

## 4.2 Best Linear Unbiased Prediction

Within the EM algorithm, the random effects,  $\gamma_i$ , are estimated by best linear unbiased prediction (BLUP).<sup>10</sup> Indeed, the E-step substitutes the  $\gamma_i$ 's by their conditional expectation given the observed data  $y_i$  and the current fit for  $\theta$ . The conditional expectation of  $\gamma_i$  given the data is

$$\begin{aligned}\hat{\gamma}_i = E(\gamma_i | y_i) &= \Delta \bar{Z}_i' \left( \bar{Z}_i \Delta \bar{Z}_i' + R_i \right)^{-1} (y_i - W_i \bar{\Gamma}) \\ &= \left( \bar{Z}_i' R_i^{-1} \bar{Z}_i + \Delta^{-1} \right)^{-1} \bar{Z}_i' R_i^{-1} (y_i - W_i \bar{\Gamma}),\end{aligned}\tag{20}$$

which is also the argument that maximizes the complete data likelihood, as defined in (9), with respect to  $\gamma_i$ . It can be noted from the first equality of (20) that this expression is analogous to the predictor of the random effects derived in Lee and Griffiths (1979), Lindley and Smith (1972) and Smith (1973). The main difference concerns the way the regression coefficients and the variances components are estimated.

The conditional variance of  $\gamma_i$  is given by

$$V_{\gamma_i} = \text{var}(\gamma_i | y_i) = \left( \bar{Z}_i' R_i^{-1} \bar{Z}_i + \Delta^{-1} \right)^{-1},\tag{21}$$

which is equivalent to the inverse of  $I(\gamma_i) = \bar{Z}_i' R_i^{-1} \bar{Z}_i + \Delta^{-1}$ , the observed Fisher information matrix obtained by taking the second derivative of the log-likelihood of the complete data with respect to  $\gamma_i$ .

These two formulae have an empirical Bayesian interpretation. Given that  $\gamma$  is random, the likelihood  $f(\gamma)$  can be thought as the “prior” density of  $\gamma$ . The posterior distribution of the latter is Normal with mean and variance given by (20) and (21), respectively.

## 4.3 E-step

At each iteration, the E-step requires the calculation of the conditional expectation of (17) given the observed data and the current fit for the parameters, to obtain an expression for  $Q_i(\theta)$ , for  $i = 1, \dots, N$ .<sup>11</sup>

To obtain  $Q_{1i}$ , we take conditional expectation of both sides of (15). Substituting

$$E_{\theta^{(b-1)}} \left( \bar{\varepsilon}_i' R_i^{-1} \bar{\varepsilon}_i | y_i \right) = \text{Tr} \left( \bar{Z}_i' R_i^{-1} \bar{Z}_i V_{\gamma_i}^{(b)} \right) + \hat{\bar{\varepsilon}}_i' R_i^{-1} \hat{\bar{\varepsilon}}_i,$$

---

<sup>10</sup>Further details are provided in Appendix A.2.

<sup>11</sup>Detailed computations are shown in Appendix A.3.

where  $\hat{\varepsilon}_i = y_i - W_i \bar{\Gamma}^{(b)} - \bar{Z}_i \hat{\gamma}_i^{(b)}$ , into  $E_{\theta^{(b-1)}} \{L_{1i} \mid y_i\}$ , yields

$$Q_{1i} = E_{\theta^{(b-1)}} (L_{1i} \mid y_i) = \mu_3 - \frac{1}{2} \log |R_i| - \frac{1}{2} \log |W_i' R_i^{-1} W_i| - \frac{1}{2} \text{Tr} \left( \bar{Z}_i' R_i^{-1} \bar{Z}_i V_{\gamma_i}^{(b)} \right) - \frac{1}{2} \hat{\varepsilon}_i' R_i^{-1} \hat{\varepsilon}_i. \quad (22)$$

where  $\hat{\gamma}_i^{(b)}$  and  $V_{\gamma_i}^{(b)}$  are given by (20) and (21) respectively, after substituting the current fit for  $\theta$  at each iteration  $b = 1, 2, \dots$

To obtain  $Q_{2i}$ , we take the conditional expectation of (16). Substituting

$$E_{\theta^{(b-1)}} (\varepsilon_i' H_i \varepsilon_i \mid y_i) = \text{Tr} \left( \bar{Z}_i' H_i \bar{Z}_i V_{\gamma_i}^{(b)} \right) + \hat{\varepsilon}_i' H_i \hat{\varepsilon}_i,$$

where  $\hat{\varepsilon}_i = y_i - W_i \bar{\Gamma} - \bar{Z}_i \hat{\gamma}_i^{(b)}$ , into  $E_{\theta^{(b-1)}} \{L_{2i} \mid y_i\}$ , yields

$$Q_{2i} = E_{\theta^{(b-1)}} (L_{2i} \mid y_i) = \mu_4 - \frac{1}{2} \log |W_i' R_i^{-1} W_i| - \frac{1}{2} \text{Tr} \left( \bar{Z}_i' H_i \bar{Z}_i V_{\gamma_i}^{(b)} \right) - \frac{1}{2} \hat{\varepsilon}_i' H_i \hat{\varepsilon}_i. \quad (23)$$

Finally, substituting

$$E_{\theta^{(b-1)}} \left( \gamma_i' \Delta^{-1} \gamma_i \mid y \right) = \text{Tr} \left( \Delta^{-1} V_{\gamma_i}^{(b)} \right) + \hat{\gamma}_i^{(b)'} \Delta^{-1} \hat{\gamma}_i^{(b)},$$

into  $E_{\theta^{(b-1)}} \{\log f(\gamma_i) \mid y_i\}$ , yields

$$Q_{3i} = E_{\theta^{(b-1)}} (\log f(\gamma_i) \mid y) = -\frac{K^*}{2} \log 2\pi + \frac{1}{2} \log |\Delta^{-1}| - \frac{1}{2} \text{Tr} \left( \Delta^{-1} V_{\gamma_i}^{(b)} \right) - \frac{1}{2} \hat{\gamma}_i^{(b)'} \Delta^{-1} \hat{\gamma}_i^{(b)}. \quad (24)$$

## 4.4 M-step

The M-Step consists in maximizing (18) with respect to the parameters of interest, contained in  $\theta$ .

**Estimation of the Average Effect.** An estimate of  $\bar{\Gamma}$  can be obtained by maximizing  $Q(\theta; \theta^{(b-1)})$  with respect to  $\bar{\Gamma}$ . This reduces to solving

$$\frac{\partial Q(\theta; \theta^{(b-1)})}{\partial \bar{\Gamma}} = \frac{\partial}{\partial \bar{\Gamma}} \left( -\frac{1}{2} \sum_{i=1}^N \hat{\varepsilon}_i' H_i \hat{\varepsilon}_i \right) = 0.$$

The solution is

$$\bar{\Gamma}^{(b)} = \left( \sum_{i=1}^N W_i' R_{i(b-1)}^{-1} W_i \right)^{-1} \sum_{i=1}^N W_i' R_{i(b-1)}^{-1} \left( y_i - \bar{Z}_i \hat{\gamma}_i^{(b)} \right). \quad (25)$$

which is equivalent to the GLS estimation of  $\bar{\Gamma}$  when the model is given by  $y_i^* = W_i \bar{\Gamma} + \varepsilon_i$ , where  $y_i^* = y_i - \bar{Z}_i \gamma_i$ , as if the  $\gamma_i$ 's were known.

**Estimation of the Variances of the Error Terms.** An estimate of  $\sigma_{\varepsilon_i}^2$  can be derived by maximizing (18). Because  $Q_{3i}$  is not a function of  $\sigma_{\varepsilon_i}^2$  and given that no information is lost by neglecting  $Q_{2i}$  (as noted by Patterson and Thompson (1971), and Harville (1977)), we base inference for  $\sigma_{\varepsilon_i}^2$  only on  $Q_{1i}$ , which is defined in (22).

Substituting  $R_i = \text{var}(\varepsilon_i) = \sigma_{\varepsilon_i}^2 I_T$  into (22) and equating the first derivative of the latter with respect to  $\sigma_{\varepsilon_i}^2$  to zero, yields

$$\sigma_{\varepsilon_i}^{2(b)} = \frac{\hat{\varepsilon}_i' \hat{\varepsilon}_i + \text{Tr}(\bar{Z}_i' \bar{Z}_i V_{\gamma_i}^{(b)})}{T - r(W_i)}, \quad (26)$$

where  $\hat{\varepsilon}_i = y_i - W_i \bar{\Gamma}^{(b)} - \bar{Z}_i \hat{\gamma}_i^{(b)}$ . A necessary condition to be satisfied is:  $T > \text{rank}(W_i)$ .

**Estimation of the Random Coefficient Variance-Covariance Matrix.** Under the law of total variance, the unconditional variance of  $\gamma_i$  can be written as

$$\begin{aligned} \Delta = \text{var}(\gamma_i) &= \text{var}[E(\gamma_i | y_i)] + E[\text{var}(\gamma_i | y_i)] \\ &= \text{var}(\hat{\gamma}_i) + E(V_{\gamma_i}). \end{aligned} \quad (27)$$

Therefore, it can be shown that

$$\hat{\Delta} = \frac{1}{N} \sum_{i=1}^N \{\hat{\gamma}_i \hat{\gamma}_i' + V_{\gamma_i}\} \quad (28)$$

is an unbiased estimator of  $\Delta$ . Indeed, taking expectation of both sides of (28) and using (27), we get

$$E(\hat{\Delta}) = \frac{1}{N} \sum_{i=1}^N \{E(\hat{\gamma}_i \hat{\gamma}_i') + E(V_{\gamma_i})\} = \frac{1}{N} \sum_{i=1}^N \{\text{var}(\hat{\gamma}_i) + E(V_{\gamma_i})\} = \Delta.$$

Notably, the EM estimator of the variance-covariance matrix of the random effects (which is the argument which maximizes (24) with respect to  $\Delta$ ) is equal to

$$\Delta^{(b)} = \frac{1}{N} \sum_{i=1}^N \{\hat{\gamma}_i^{(b)} \hat{\gamma}_i^{(b)'} + V_{\gamma_i}^{(b)}\}, \quad (29)$$

which is equivalent to (28) after substituting the unknown parameters with their current fit in the EM algorithm.<sup>12</sup>

---

<sup>12</sup>See Appendix A.4 for computations.

## 4.5 EM Algorithm: Complete Iterations

The EM algorithm steps can be summarised as follows. We start with some initial guess:  $\psi^{(0)}$ ,  $\Delta_{(0)}$  and  $R_{i(0)} = \sigma_{\varepsilon_i}^{2(0)} I_{T-p}$ . We suggest using Swamy (1970) estimates, which are reported in the next Section, since they are consistent estimators of the average effects and the variance components. Then, for  $b = 1, 2, \dots$

1. Given the current fit for  $\theta$  at iteration  $b$ , we compute  $\text{var}(\gamma_i \mid y_i, \theta^{(b-1)})$  and  $E_{\theta^{(b-1)}}(\gamma_i \mid y_i)$ , which are given by

$$\begin{aligned} V_{\gamma_i}^{(b)} &= \left( \bar{Z}_i' R_{i(b-1)}^{-1} \bar{Z}_i + \Delta_{(b-1)}^{-1} \right)^{-1}, \\ \hat{\gamma}_i^{(b)} &= V_{\gamma_i}^{(b)} \bar{Z}_i' R_{i(b-1)}^{-1} (y_i - W_i \bar{\Gamma}^{(b-1)}), \end{aligned}$$

respectively.

2. The average coefficients are given by

$$\bar{\Gamma}^{(b)} = \left( \sum_{i=1}^N W_i' R_{i(b-1)}^{-1} W_i \right)^{-1} \sum_{i=1}^N W_i' R_{i(b-1)}^{-1} (y_i - \bar{Z}_i \hat{\gamma}_i^{(b)}).$$

3. Finally, we can compute, the variance components:

$$\sigma_{\varepsilon_i}^{2(b)} = \frac{\hat{\varepsilon}_i \hat{\varepsilon}_i' + \text{Tr}(\bar{Z}_i' \bar{Z}_i V_{\gamma_i}^{(b)})}{T - r(W_i)},$$

where  $\hat{\varepsilon}_i = y_i - W_i \bar{\Gamma}^{(b)} - \bar{Z}_i \hat{\gamma}_i^{(b)}$  and

$$\Delta^{(b)} = \frac{1}{N} \sum_{i=1}^N \{ V_{\gamma_i}^{(b)} + \hat{\gamma}_i^{(b)} \hat{\gamma}_i^{(b)'} \}.$$

The iterations continue until the difference  $L(y; \theta^{(b)}) - L(y; \theta^{(b-1)})$  changes only by an arbitrary small amount, where  $L(y; \theta)$  is the likelihood of the observed data.

## 5 Comparison between EM-REML Estimation and Alternative Methods.

In this section, we review some of the existing sampling and Bayesian methods commonly used to estimate heterogeneous panel data, to highlight similarities and differences with the EM-REML approach.



## 5.1 Average Effects

Following Searle (1978, eq. 3.17), representing (25) and (20) as a system of two equations, we can rewrite these two formulae as

$$\hat{\bar{\Gamma}} = \left( \sum_{i=1}^N W_i' V_i^{-1} W_i \right)^{-1} \sum_{i=1}^N W_i' V_i^{-1} y_i, \quad (30)$$

$$\hat{\gamma}_i = \Delta \bar{Z}_i' V_i^{-1} \left( y_i - W_i \hat{\bar{\Gamma}} \right), \quad (31)$$

respectively. Note that  $\hat{\bar{\Gamma}}$  is the estimator which maximizes the log-likelihood function constructed by referring to the marginal distribution of the dependent variable. When  $f_i = 1$  for all  $i$ , and  $W_i = \bar{Z}_i$ , equation (30) is related to the Swamy GLS estimator. The latter can be rewritten as a weighted average of the least squares estimates of the individual units:

$$\hat{\bar{\Gamma}} = \sum_{i=1}^N \Psi_i \hat{\psi}_{i,ols}, \quad (32)$$

where

$$\begin{aligned} \Psi_i &= \left\{ \sum_{i=1}^N [\Delta + \sigma_{\varepsilon_i}^2 (\bar{Z}_i' \bar{Z}_i)^{-1}]^{-1} \right\}^{-1} [\Delta + \sigma_{\varepsilon_i}^2 (\bar{Z}_i' \bar{Z}_i)^{-1}]^{-1}, \\ \hat{\psi}_{i,ols} &= (\bar{Z}_i' \bar{Z}_i)^{-1} \bar{Z}_i' y_i. \end{aligned} \quad (33)$$

Swamy's estimator is a two-step procedure, which requires first to estimate  $N$  time series separately as if the individual coefficients were fixed (in the sense that they are not realizations from a common distribution) and all different in each cross-section. Instead, the EM-REML is an iterative method which shrinks the unit-specific parameters towards a common mean. Maddala et al. (1997) argue in favour of iterative procedures when the model includes lagged dependent variables since, as indicated in Amemiya and Fuller (1967), Maddala (1971) and Pagan (1986), when estimating dynamic models, the two-step estimators based on any consistent estimators of  $\sigma_{\varepsilon_i}^2$  and  $\Delta$  are consistent but not efficient.

Hsiao, Pesaran and Tahmiscioglu (1999) show that  $\hat{\bar{\Gamma}}$  is equivalent to the posterior mean of  $\bar{\Gamma}$  in a Bayesian approach which assumes the prior distribution of  $\bar{\Gamma}$  is normal with mean  $\mu$  and variance  $\Omega$ , with  $\Omega^{-1} = 0$ . Another important contribution of the aforementioned paper is to establish that the Bayes estimator  $\hat{\bar{\Gamma}}$  is asymptotically equivalent to the mean group estimator proposed by Pesaran and Smith (1995), as  $T \rightarrow \infty$ ,  $N \rightarrow \infty$ , and  $\sqrt{N}/T \rightarrow 0$ .

## 5.2 Unit-Specific Parameters

Without loss of generality, for comparison purposes, let us focus on the case where  $f_{1i} = 1$ ,  $\forall i$  and  $\underline{Z}_i = 0$ . Substituting (30) and (20) into (5) yields the best linear unbiased predictor

of  $\psi_i$  which following Lee and Griffiths (1979), can be rewritten as

$$\begin{aligned}\hat{\psi}_i &= \hat{\bar{\Gamma}} + \hat{\gamma}_i \\ &= \left(\bar{Z}_i' R_i^{-1} \bar{Z}_i + \Delta^{-1}\right)^{-1} \left(\left(\bar{Z}_i' R_i^{-1} \bar{Z}_i\right) \hat{\psi}_{i,ols} + \Delta^{-1} \hat{\bar{\Gamma}}\right).\end{aligned}\tag{34}$$

The latter expression is also related to the empirical Bayes estimator of  $\psi_i$ , described in Maddala et al. (1997). The EM-REML predictor of  $\psi_i$  is thus a weighted average between the OLS estimator of  $\psi_i$  and the estimator of the overall mean,  $\bar{\Gamma}$ , given by (30). Interestingly, as shown in Smith (1973), the latter can be rewritten as a simple average of the  $\hat{\psi}_i$ :

$$\hat{\bar{\Gamma}} = \frac{1}{N} \sum_{i=1}^N \hat{\psi}_i.\tag{35}$$

**Mean Group and Shrinkage Estimators.** When the time dimension is large enough (relative to the number of parameters to be estimated), it is sensible to estimate a different time-series model for each unit, as proposed by Pesaran and Smith (1995). Besides its simplicity, one strong advantage of their Mean Group (MG) estimator is that it does not require to impose any assumption on the distribution of the unit-specific coefficients. However, a drawback of the MG estimation is that it may perform rather poorly when either  $N$  or  $T$  are small (Hsiao, Pesaran and Tahmiscioglu, 1999). Moreover, as noted in Smith and Fuertes (2016), the MG estimator is very sensitive to outliers. Boyd and Smith (2002) find that the weighting which the Swamy estimator applies, may not suffice to reduce this problem. To overcome the latter, one could either consider robust versions which trim the outliers to minimize their effect, or shrinkage methods. Maddala et al. (1997), estimating short-run and long-run elasticities of residential demand for electricity and natural gas, find that individual heterogeneous state estimates are difficult to interpret and have the wrong signs. They suggest shrinkage estimators (instead of heterogeneous or homogeneous parameter estimates) if one is interested in obtaining elasticity estimates for each state since these give more reliable results. Our estimation method belongs to the class of shrinkage estimators. In fact, the unobserved idiosyncratic components of the random coefficients,  $\gamma_i$ , are estimated by BLUP. This choice arises naturally in the EM algorithm, and in some applications may be advantageous compared to estimating  $N$  time series separately since BLUP estimates tend to be closer to zero than the estimated effects would be if they were computed by treating a random coefficient as if it were fixed. Shrinkage approaches can be seen as an intermediate strategy between heterogeneous models (which avoid bias) and pooled methods (which allow for efficiency gains), and therefore might help reducing the trade-off between bias and efficiency discussed in Baltagi, Bresson and Pirotte (2008). As shown in the Monte Carlo analysis, as  $T \rightarrow \infty$  the difference between the Swamy, the MG, and the EM-REML estimators goes to zero. Finally, our approach can be advantageous (*i*) when individual-specific

characteristics which do not vary over time enter the regression equation, and (ii) when the interest lies in explaining the drivers of coefficients heterogeneity. In the first case, computing the OLS estimates for each unit is not feasible. In the second case, if  $N$  is large one could first estimate  $N$  time series separately and in a second step regress the OLS estimates on a set of unit-specific characteristics. Instead, our likelihood approach does not require  $N$  to be very large.

### 5.3 Variance Components

We now compare the EM-REML estimator of the random coefficient variance-covariance matrix, given by (29), with the Swamy (1970) and Lee and Griffiths (1979) estimators. Swamy suggested estimating  $\text{var}(\gamma_i)$  as

$$\hat{\Delta}_S = \hat{\Delta}_{S_1} - N^{-1} \sum_{i=1}^N \text{var}(\hat{\psi}_{i,ols}), \quad (36)$$

where

$$\hat{\Delta}_{S_1} = \frac{1}{N-1} \sum_{i=1}^N \left( \hat{\psi}_{i,ols} - N^{-1} \sum_{i=1}^N \hat{\psi}_{i,ols} \right) \left( \hat{\psi}_{i,ols} - N^{-1} \sum_{i=1}^N \hat{\psi}_{i,ols} \right)', \quad (37)$$

$\hat{\psi}_{i,ols}$  are obtained by estimating  $N$  time series separately by OLS,  $\text{var}(\hat{\psi}_{i,ols}) = \hat{\sigma}_{\varepsilon_i}^2 (\bar{Z}_i' \bar{Z}_i)^{-1}$ , and

$$\hat{\sigma}_{\varepsilon_i}^2 = \frac{1}{T - K^*} (y_i - \bar{Z}_i' \hat{\psi}_{i,ols})' (y_i - \bar{Z}_i' \hat{\psi}_{i,ols}) \quad (38)$$

are the OLS estimated variances of the error terms. However, (36) is not necessarily non-negative definite. Therefore, if that is the case the author suggests considering only (37). The latter estimator is nonnegative definite and consistent when  $T$  tends to infinity. This estimator is also used in the empirical Bayesian approach and in Lee and Griffiths' "modified mixed estimation" procedure. Unfortunately, this estimator can be severely biased in finite sample. Another drawback of (36) is that it is subject to large discontinuities.<sup>13</sup> As shown in the Monte Carlo analysis, the root mean square errors of this estimator can be quite large. To understand, note that the estimator to be used in practical applications can be rewritten as

$$\hat{\hat{\Delta}} = \mathbb{I}(\hat{\Delta} > 0) \hat{\Delta} + \mathbb{I}(\hat{\Delta} \leq 0) \hat{\Delta}_{S_1},$$

where  $\mathbb{I}(A) = 1$  if event  $A$  occurs. Focusing on the  $k$ th diagonal element, and assuming for illustrative purposes that

$$\hat{\Delta}_{S_1,k} = 2, \quad \text{var}(\hat{\psi}_{ik}) = \left\{ N^{-1} \sum_{i=1}^N \text{var}(\hat{\psi}_{ik}) \right\} \in \{1, 2, 3, 4\},$$

---

<sup>13</sup>I am grateful to Ron Smith who pointed out this issue in a meeting.

we have

$$\hat{\hat{\Delta}}_k = \begin{cases} 2 & \text{if } v\hat{a}r(\hat{\psi}_{ik}) \in \{2, 3, 4\} \\ 1 & \text{if } v\hat{a}r(\hat{\psi}_{ik}) = 1 \end{cases}$$

When the variances are unknown, Lee and Griffiths (1979) suggest maximizing the joint likelihood of the random coefficients and the observed data given in (9) with respect to the unknown parameters of the model, to get the following iterative solutions of the variance components:<sup>14</sup>

$$\hat{\sigma}_{\varepsilon_i}^2 = \frac{1}{T} (y_i - \bar{Z}_i \hat{\psi}_i)' (y_i - \bar{Z}_i \hat{\psi}_i), \quad (39)$$

where  $\hat{\psi}_i$  is given by (34), and

$$\hat{\Delta}_{LG} = \frac{1}{N} \sum_{i=1}^N \hat{\gamma}_i \hat{\gamma}_i'. \quad (40)$$

Within the EM algorithm, the random effects,  $\gamma_i$ , are considered as missing data and replaced by their conditional expectation given the data, which yields the BLUP of  $\gamma_i$ . At the same time, we have seen that the latter is equivalent to the argument which maximizes the joint likelihood of the observed data and random effects, given in (9). This is the approach followed by Lee and Griffiths (1979). We argue in favor of treating the joint likelihood as an incomplete data problem to then applying the EM algorithm to obtain maximum likelihood estimates because, among the other reasons highlighted in Section 4, the estimator given by (40) does not satisfy the law of total variance. This is not the case when applying the EM algorithm. Consequently, our approach has an advantage over both Swamy (1970) and Lee and Griffiths (1979) in finite sample, since

$$E(\hat{\Delta}_{LG}) \leq E(\hat{\Delta}_{EM}) \equiv \Delta \leq E(\hat{\Delta}_{S_1}), \quad (41)$$

where

$$\hat{\Delta}_{EM} = \frac{1}{N} \sum_{i=1}^N \{V_{\gamma_i} + \hat{\gamma}_i \hat{\gamma}_i'\} \quad (42)$$

is the maximum likelihood estimator obtained by applying the EM algorithm. Result (41) is of relevance because  $\Delta$  appears not only in both the formula for the average effect and the predicted random effects but also in their standard errors. Testing hypothesis crucially depends on correctly estimating the random coefficient variances.

Finally, we report the Bayes mode of the posterior distribution of  $\Delta$  and  $\sigma_{\varepsilon_i}^2$  suggested by Lindley and Smith (1972) and Smith (1973), which are equal to

$$\hat{\sigma}_{\varepsilon_i}^2 = \frac{1}{T + v_i + 2} \left\{ v_i \lambda_i + (y_i - \bar{Z}_i \hat{\psi}_i)' (y_i - \bar{Z}_i \hat{\psi}_i) \right\}, \quad (43)$$

---

<sup>14</sup>In this Section, we omit the superscript  $b = 1, 2, \dots$  in  $\hat{\psi}_i^{(b)}$  and  $\hat{\gamma}_i^{(b)}$  for ease of exposition even though the solutions are iterative.

$$\bar{\Delta} = \frac{1}{N + \rho - K^* - 2} \left\{ \Upsilon + \sum_{i=1}^N \hat{\gamma}_i \hat{\gamma}_i' \right\}, \quad (44)$$

respectively, under the assumption that  $\Delta^{-1}$  has a Wishart distribution, with  $\rho$  degrees of freedom and matrix  $\Upsilon$  and  $\sigma_{\varepsilon_i}^2$  follows a  $\chi^2$  with prior parameters  $v_i$  and  $\lambda_i$ , and is independent of  $\Delta$ . Note from (34) that  $\hat{\gamma}_i = \hat{\psi}_i - \hat{\Gamma}$ . Smith (1973) suggests vague priors by setting  $\rho = 1$  and  $\Upsilon$  to be a diagonal matrix with small positive entries (such as .001). We note that, by setting  $\rho = K^* + 2$ ,  $v_i = -r(W_i) - 2$  and  $v_i \lambda_i = Tr(\bar{Z}_i' \bar{Z}_i \Upsilon)$ , we can draw an analogy between the EM-REML estimates, given by (42) and (26), and the modes of the posterior distributions of  $\Delta$  and  $\sigma_{\varepsilon_i}^2$ , given by (44) and (43), respectively.

## 5.4 Comparison between EM and a Full Bayesian Implementation

We can now compare the EM approach to the Bayesian estimation. The EM algorithm gives a probability distribution over the random effects,  $\gamma$ , together with a point estimate for  $\theta$ , the vector of average coefficients and variance components of the model. The latter is treated as being random in a full Bayesian version. The advantage of the EM compared to the iterative Bayesian approach developed by Lindley and Smith (1992) and the Gibbs sampling-based approach suggested in Hsiao, Pesaran and Tahmiscioglu (1999), would be that there is no need to specify prior means and variances, the choice of which may not be always obvious. At the same time, as discussed in Kass and Wasserman (1996), when sample sizes are small (relative to the number of parameters being estimated) the prior choice will have a heavy weight on the posterior, which will consequently be far from being data dominated. While the Bayesian point estimates incorporate prior information, the EM-REML estimates do not involve the starting values (chosen to initiate the algorithm). One can start with any initial value. As shown in Dempster et al. (1977), the incomplete-data likelihood function  $L(y; \theta)$  does not decrease after an EM iteration, that is  $L(y; \theta^{(b)}) \geq L(y; \theta^{(b-1)})$  for  $b = 1, 2, \dots$ . Nevertheless, this property does not guarantee convergence of the EM algorithm since it can get trapped in a local maximum. In complex cases, Pawitan (2001) suggests to try several starting values or to start with a sensible estimate. However, in the context of random coefficient models the choice of Swamy (1970) estimates as starting values is rather natural, as they are consistent parameter estimates.

Moreover, using a purely “noninformative” prior (in the sense of Koop (2003)) may have the undesirable property that this prior “density” does not integrate to one, which in turn may raise many of the problems discussed in the Bayesian literature (e.g. Hobert and Casella (1996)). For instance, assuming that  $\Delta^{-1}$  has a Wishart distribution with scale matrix  $(\rho \Upsilon)$  and  $\rho$  degrees of freedom, Hsiao, Pesaran and Tahmiscioglu (1999) note that the bias of

both the empirical and hierarchical Bayes estimators of the regression coefficients is sensitive to the specification of the prior scale matrix. Being unable to use a diffuse prior for the covariance matrix, which would cause their Gibbs algorithm to break down, they set  $\Upsilon = \hat{\Delta}_S$ , the Swamy estimator of the random coefficient covariance matrix. If the latter is negative definite, the consistent (but biased) version (37) must be used, affecting the Bayes estimates of the regression coefficients adversely.

Finally, it is known that the EM algorithm may converge slowly. However, in the context of random coefficient models, convergence is usually achieved almost as quickly as in the Gibbs sampler.<sup>15</sup>

## 6 Hypothesis Testing

### 6.1 Inference for Fixed Coefficients

**Covariance Matrix of the Estimator of the Fixed Coefficients.** Unlike the Newton-Raphson and related methods, the EM algorithm does not automatically provide an estimate of the covariance matrix of the maximum likelihood estimates. However, in the context of the random coefficient type models here considered, the Fisher information matrix  $I(\bar{\Gamma}^{(B)})$  can be easily derived by evaluating analytically the second-order derivatives of the marginal log-likelihood of the observed data ( $\log f(y; \theta)$ ) since computations are not complicated. Therefore, after convergence, the standard errors of  $\bar{\Gamma}^{(B)}$  can be computed as the square root of the diagonal elements of the inverse of the Fisher information matrix, given by

$$\hat{\Phi} = \left( \sum_{i=1}^N W_i' V_{i(B)}^{-1} W_i \right)^{-1}, \quad (45)$$

where  $V_i = \text{var}(y_i) = \bar{Z}_i \bar{\Delta} \bar{Z}_i' + R_i$ , while  $B$  denotes the last iteration of the EM algorithm.

**Adjusted Estimator of the Covariance Matrix of Fixed Coefficients.** Let  $\tilde{\bar{\Gamma}} = \bar{\Gamma}^{(B)}$  be the “feasible” estimator of  $\bar{\Gamma}$  obtained by substituting the unknown parameters with their estimates into the “infeasible” estimator  $\hat{\bar{\Gamma}}$ , given by equation (30). We define  $\Phi = \text{var}(\hat{\bar{\Gamma}})$ ,

---

<sup>15</sup>For instance, in the panel model used in the application, with  $N = 38$  and  $60 \leq T_i \leq 87$ , and  $K = 8$  regressors, including the constant, the EM algorithm converges after around 17 seconds. The Gibbs sampling algorithm is quicker, requiring around 10 seconds to run 5000 iterations. If we increase the number of regressors to 20, the difference slightly increases, with the EM algorithm and the Gibbs sampler requiring around 40 and 15 seconds, respectively. Despite being slower than its Bayesian counterpart, the EM algorithm converges rather quickly.

which is a function of  $\vartheta = (\omega', \sigma_\varepsilon^2)'$ , the  $\bar{r} \times 1$  vector of variance-covariance parameters of the model.

We note that  $\hat{\Phi} = \Phi(\hat{\vartheta})$  is a biased estimator of  $\text{var}(\tilde{\tilde{\Gamma}})$ . The literature on linear mixed models offers good insights into the two main sources of this bias. First,  $\Phi(\vartheta)$  takes no account of the variability of  $\hat{\vartheta}$  in  $\tilde{\tilde{\Gamma}}$ . This problem was addressed by Kackar and Harville (1984). Second,  $\hat{\Phi}$  underestimates  $\Phi$ , as shown by Kenward and Roger (1997). The solution provided by the latter can be easily applied into our setting to obtain an estimator of  $\text{var}(\tilde{\tilde{\Gamma}})$ ,  $\hat{\Phi}_A$ , which incorporates the necessary adjustments to correct both form of bias.<sup>16</sup>

**Hypothesis Testing of Average Effects.** To test the hypothesis  $\bar{\Gamma} = \bar{\Gamma}_0$ , for  $\bar{\Gamma}_0$  a known  $\bar{K} \times 1$  vector, we use the following criterion suggested by Swamy (1970):

$$\frac{N - \bar{K}}{\bar{K}(N - 1)} (\tilde{\tilde{\Gamma}} - \bar{\Gamma})' \hat{\Phi}_A^{-1} (\tilde{\tilde{\Gamma}} - \bar{\Gamma}), \quad (46)$$

whose asymptotic distribution is F, with  $\bar{K}$ ,  $N - \bar{K}$  degrees of freedom.

## 6.2 Assessing the Precision for the Unit-Specific Coefficients

In the general case, the standard errors of the predictor of  $\psi_{1i}$  can be computed as the square root of the diagonal elements of

$$\text{var}(\hat{\psi}_{1i} - \psi_{1i}) = F_{1i} \Phi F_{1i}' + \text{var}(\hat{\gamma}_i - \gamma_i) - F_{1i} \Lambda - \Lambda' F_{1i}', \quad (47)$$

where

$$\begin{aligned} \Lambda &= \text{cov}(\hat{\tilde{\tilde{\Gamma}}} - \bar{\Gamma}, \gamma_i) = \Phi W_i' V_i^{-1} \bar{Z}_i \Delta, \\ \text{var}(\hat{\gamma}_i - \gamma_i) &= \Delta \left[ I - \bar{Z}_i' V_i^{-1} (I + W_i \Phi W_i' V_i^{-1}) \bar{Z}_i \Delta \right], \end{aligned}$$

and  $\Phi = \text{var}(\hat{\tilde{\tilde{\Gamma}}})$  as defined in (45).<sup>17</sup>

At the same time, one can exploit the fact that the EM algorithm provides a distribution over the random effects. For instance, we suggest drawing  $S$  samples from

$$\gamma_i^{(s)} \mid y_i \sim N(\hat{\gamma}_i, V_{\gamma_i}), \quad (48)$$

<sup>16</sup>Some details for computation are given in Appendix A.5.1. A Matlab code to obtain  $\hat{\Phi}_A$  is provided.

<sup>17</sup>Expression (47) is equivalent to the one proposed by Lee and Griffiths (1979). See Appendix A.5.2 for further details. The estimator of  $\psi_i$  derived in equation (34) has been obtained under the assumption that  $F_{1i} = I$ ,  $\forall i$ . Therefore, its standard errors can also be obtained from (47) after substituting  $F_{1i} = I$ ,  $\forall i$ .

where  $\hat{\gamma}_i$  and  $V_{\gamma_i}$  are given by (20) and (21) respectively, to then report histograms for each unit for comparison and diagnostic purposes.

## 7 Monte Carlo Simulations

In this section, we employ Monte-Carlo experiments to examine and compare the finite sample properties of the proposed EM-REML method, the Swamy's random coefficient model, and the Mean Group (MG) estimation. We report results on the bias and root mean square error (RMSE) of the average effects and of the variance components of the model. Particular attention is also paid to the accuracy of the estimated standard errors and to the power performances of the estimators.

### 7.1 Data Generating Process

The data generating process (DGP) used in the Monte Carlo analysis is given by

$$\begin{aligned} y_{it} &= c_i + \beta_i x_{it} + \phi_i y_{it-1} + \varepsilon_{it}, \\ x_{it} &= c_{x,i}(1 - \rho) + \rho x_{it-1} + u_{it}, \end{aligned} \tag{49}$$

where

$$\begin{aligned} \varepsilon_{it} &\sim i.i.d.N(0, \sigma_{\varepsilon_i}^2), \\ u_{it} &\sim i.i.d.N(0, 1), \\ c_{x,i} &\sim i.i.d.N(1, 1). \end{aligned} \tag{50}$$

The sample sizes considered are  $N = \{30, 50\}$  and  $T = \{10, 20, 30, 40, 50, 60, 80, 100\}$ . We set  $\rho = 0.6$ . Once generated, the  $x_{it}$  are taken as fixed across different replications. The variances of the time-varying disturbances are generated from  $\sigma_{\varepsilon_i}^2 = (\zeta \bar{x}_i)^2$ , where  $\bar{x}_i = T^{-1} \sum_{t=1}^T x_{it}$ , and  $\zeta = 0.5$ . The coefficients differ randomly across units according to

$$\begin{aligned} c_i &= c + \gamma_{1i}, \\ \beta_i &= \beta + \gamma_{2i}, \\ \phi_i &= \phi + \gamma_{3i}, \end{aligned} \tag{51}$$

where  $\psi = (c, \beta, \phi) = (0, 0.1, 0.5)$ . Moreover, we assume that  $\gamma_{ji} \sim i.i.d.N(0, \sigma_{\gamma_j}^2)$ , for  $j = 1, 2, 3$ . We set  $\sigma_{\gamma_1} = 0.1$ , and  $\sigma_{\gamma_2} = 0.224$ . We choose  $\sigma_{\gamma_3} = 0.07$  in order to avoid explosive behaviour. Under these settings the median signal-to-noise ratio corresponding to the slope parameters ( $\sigma_{\gamma_2}^2 / \sigma_{\varepsilon_i}^2$ ) for  $N = 30$  and averaged across the different  $T$  cases, is equal to 0.1950.<sup>18</sup>

---

<sup>18</sup>The cross-section average, computed as  $N^{-1} \sum_{i=1}^N \sigma_{\gamma_2}^2 / \sigma_{\varepsilon_i}^2$ , and averaged across the different  $T$  cases, is higher and equal to 10.63, partly due to the fact that some of the draws of  $\sigma_{\varepsilon_i}^2$  are smaller than  $\sigma_{\gamma_2}^2$ .



The initial values for the dependent variables are generated from

$$y_{i0} = \bar{\theta}_{i0} + v_{i0},$$

for  $i = 1, \dots, N$ , where  $v_{i0} \sim N(0, \sigma_v^2)$ , and

$$\begin{aligned} \bar{\theta}_{i0} &= E(y_{i0} | \gamma_i) = \sum_{s=0}^{\infty} \phi_i^s x_{i,-s} \beta_i + \frac{c_i}{1-\phi_i}, \\ \sigma_v^2 &= \text{var}(y_{i0} | \gamma_i) = \text{var}\left\{\sum_{s=0}^{\infty} \phi_i^s \varepsilon_{i,-s}\right\} \\ &= \frac{\sigma_{\varepsilon_i}^2}{1-\phi_i^2}. \end{aligned}$$

In practice, we consider only a finite number of  $x_{i,-s}$ . For each  $i$ , we generate 10 observations  $(x_{i0}, \dots, x_{i,-9})$  given that when  $|\phi_i| < 1$ , the contribution of earlier observations is quite low. The vector  $(x_{i0}, \dots, x_{i,-9})$  is not used for estimation and inference.

## 7.2 Monte Carlo Results

In this subsection, we describe the results based on 500 replications. Table 3 reports the bias and the root mean square errors (RMSE) of the EM-REML estimators of the average effects and of the variances of the random coefficients, as well as the standard errors of such biases, for  $N = 30$  and  $T = \{10, 20, 30, 40, 50, 60, 80, 100\}$ .<sup>19</sup> An overall measure of the bias of the estimated average coefficients (which is chosen to be the Euclidean norm of the bias of  $\psi$ ), and two measures of the accuracy of the estimated standard errors are also given. Table 4, and 5 describe the results for Swamy (1970), and the MG estimator, respectively.

Using the data simulated from the DGP described in the previous subsection, we find that the EM-REML approach does quite well even when the sample size is not too large. In many cases, it outperforms both Swamy and the MG estimator in term of bias of both the average effects and the variance components. For any time dimension, the REML estimators of the average coefficients and the variance components obtained applying the EM algorithm have smaller RMSE than the MG one. The RMSE of the EM-REML estimators are also smaller than the Swamy one, unless  $T$  is quite large, in which case they almost coincide.

The bias of the EM-REML estimator of the common intercept is equal to 0.0015 when  $T = 10$ , and to 0.0005 when  $T = 20$ . When  $T = 100$ , the bias amounts to  $-0.0007$ . In most of the cases, it is smaller than the bias of Swamy and the MG estimators, and it has lower RMSE.

---

<sup>19</sup>Similar results hold for  $N = 50$ , which we do not report here.

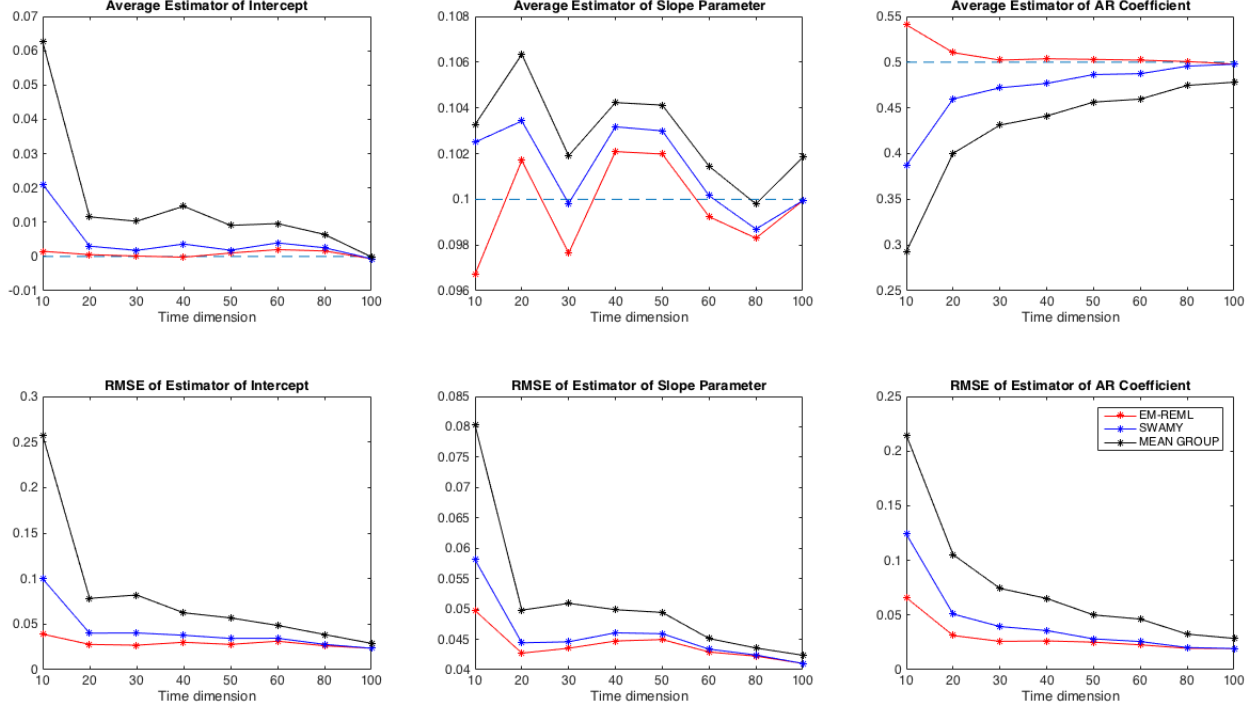


Figure 1: **Upper panel:** The estimators of the intercept (left), slope (middle) and autoregressive parameter (right panel), averaged across the 500 replications, are plotted for  $N = 30$  and  $T = \{10, 20, 30, 40, 50, 60, 80, 100\}$ . The dashed blue lines indicate the true values (used to simulate the data). The red, blue, and black solid lines correspond to the EM-REML, Swamy, and Mean Group estimator, respectively. The distances between those lines and the one corresponding to the true value measure the bias of the estimators. **Lower panel:** the root mean square errors (RMSE) of the estimators are reported.

Regarding the slope coefficient associated to  $x_{it}$ , the bias of the EM-REML estimator is equal to  $-0.0033$  when  $T = 10$ , which amounts to  $-3.3$  percent of the true value. When  $T = 20$ , the bias reduces to 1.7 percent of the true value till becoming equal to 0.1 percent when  $T = 100$ . In some cases, the EM-REML estimator may have a slightly larger bias than the Swamy one but in all cases it has a smaller or at most equal RMSE.

The advantages in term of bias of the EM-REML approach are even more notable when considering the autoregressive coefficient. For instance, when  $T = 10$ , the bias of the EM-REML estimator is equal to 0.0408, which is equivalent to 8.16% of the true value. The biases of Swamy GLS and the MG estimators of the autoregressive coefficient, when  $T = 10$ , are larger and equal to  $-22.6\%$  and  $-41.44\%$  of the true value, respectively. As expected, the bias reduces as  $T$  increases. Monte Carlo experiments corroborate Maddala et al. (1997) argument in favor of iterative procedures to two-step estimators when the model is dynamic and confirm Hsiao, Pesaran and Tahmiscioglu (1999) finding that the MG estimator is unlikely to be an

appropriate estimator when either  $N$  or  $T$  are small.

A graphical summary of these results is provided in Figure 1. The upper panels show the average values (across 500 Monte Carlo replications) of the EM-REML, Swamy, and MG estimators of the average effects. The differences between the latter and the corresponding true values measure the bias of the estimates. The RMSE of the estimators are depicted in the lower panels.

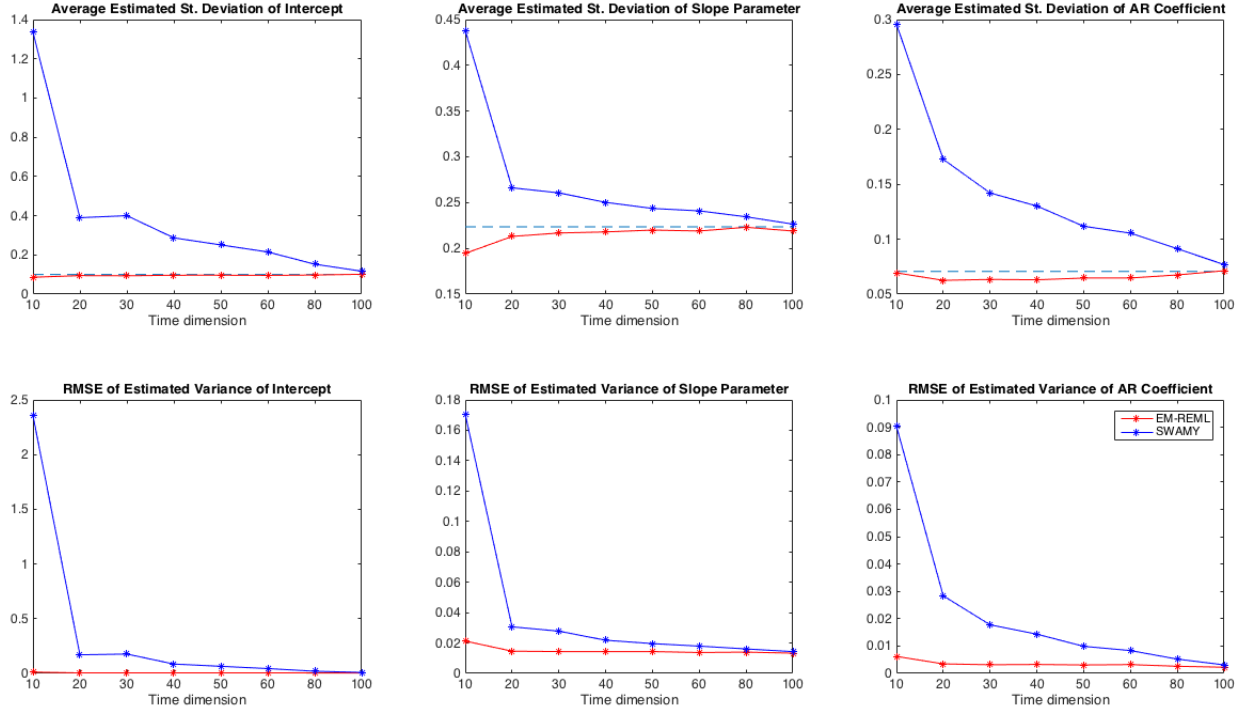


Figure 2: **Upper panel:** The estimated standard deviation of the intercept (left), slope (middle) and autoregressive parameter (right panel), averaged across the 500 replications, are plotted for  $N = 30$  and  $T = \{10, 20, 30, 40, 50, 60, 80, 100\}$ . The dashed blue line indicates the true value (used to simulate the data). The red, and blue lines correspond to the EM-REML, and Swamy estimator, respectively. The distances between those lines and the one corresponding to the true value measure the bias of the estimators. **Lower panel:** the root mean square errors (RMSE) of the estimated variances are reported.

Figure 2 illustrates the performance of the EM-REML and Swamy estimators of the random coefficients' variances. As expected, the latter largely overestimates the true variance components of the model. The size of the bias can be substantial unless the time dimension is quite large. For example, when  $T = 10$ , the probability that the Swamy unbiased estimator of the covariance matrix is negative definite is equal to 81 percent. This means that in most of the cases it has to be replaced by its consistent but biased version. At the same time,

given that in some replications we are able to use the unbiased estimator and in others only the consistent one, the RMSE of the estimated variance components obtained using Swamy procedure can be quite substantial although it reduces as  $T$  increases.

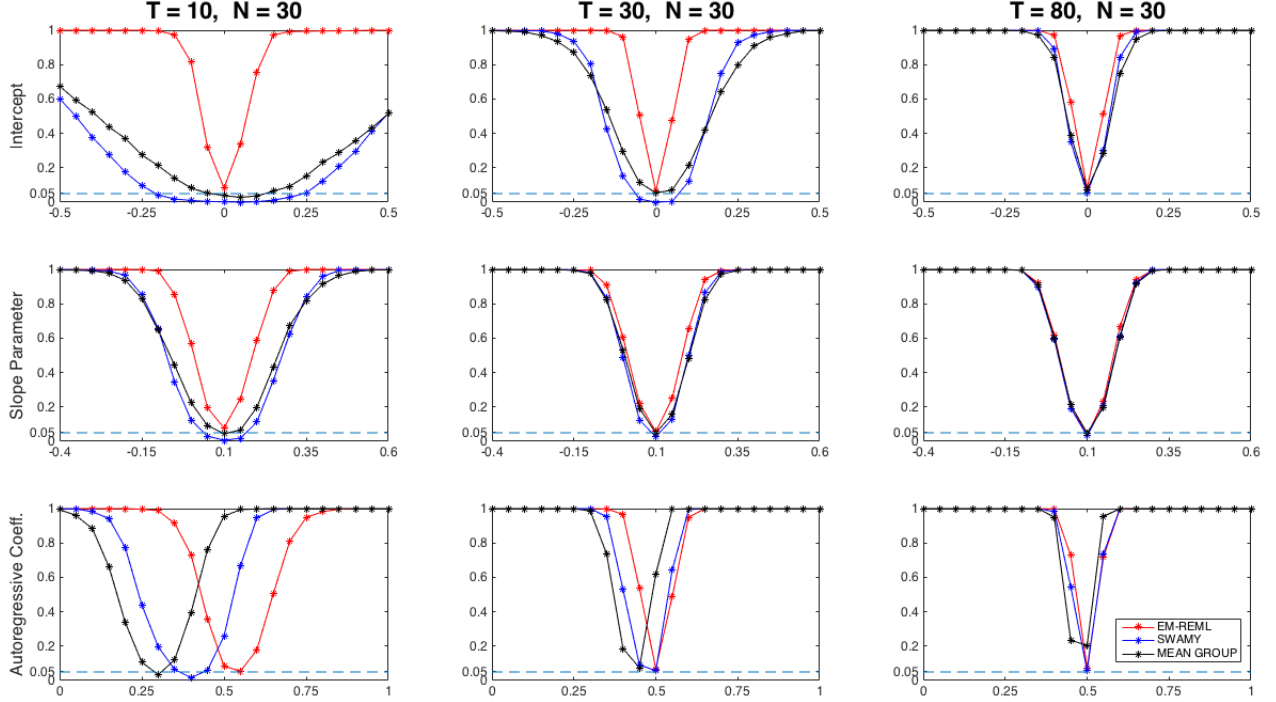


Figure 3: Rejection frequency at 5% nominal size, for the intercept (upper panels), the slope (middle panels), and autoregressive parameters (lower panels), when  $(c, \beta, \phi) = (0, 0.1, 0.5)$ . The panels on the left show results for  $(T, N) = (10, 30)$ , the panels on the middle for  $(T, N) = (30, 30)$ , and those on the right for  $(T, N) = (80, 30)$ . The red, blue, and black lines denote the power performances of the EM-REML, Swamy, and Mean Group estimators respectively.

To examine the consequences of overestimating or underestimating the true random coefficient variances when testing hypotheses, we consider the ratio between the “infeasible” standard errors (which are obtained substituting the true values used to generate the DGP into equation (45)) and the estimated standard errors of the average effects. Another important measure for inference is the accuracy of the estimated standard errors as approximations to the correct sampling standard deviation of the estimator of interest.<sup>20</sup> These ratios should ideally be equal to one. Results are reported in Tables 3, 4, and 5. We find that the standard

<sup>20</sup>In particular, the accuracy of the estimated standard errors is computed as the ratio of the latter averaged across  $B = 500$  replications,  $B^{-1} \sum_{b=1}^B \left\{ \sqrt{\hat{v} \hat{r} \left( \hat{\psi}_{k,(b)} \right)} \right\}$ , and the sampling standard deviation of the

errors obtained estimating the parameters of the model using Swamy GLS approach, are in many cases largely overestimated unless  $T$  is quite large. In the latter case, the percentage of replications in which the Swamy estimator of the random coefficient covariance matrix is negative definite diminishes, and the ratio of standard errors approaches one.

Our Monte Carlo experiments reveal that the biases of the Swamy estimator of the variance components and of the resulting standard errors can be too large to be neglected. This in turn affects hypothesis tests adversely. To demonstrate the latter point, we consider the power performances of the various estimators. We plot the power functions in Figure 3. They are computed using the Swamy type test described in equation (46) for  $N = 30$ , and various  $T$ .<sup>21</sup> It is shown that the EM-REML approach performs comparatively well even when the sample size is small. When  $T$  is small the power functions of the Swamy and MG estimators of the autoregressive coefficients are not centred at the true value of  $\phi = 0.5$ . As the time dimension increases the differences in the power performances of the various estimators reduce.

### 7.2.1 The Sensitivity of the Bayesian Estimator to the Choice of the Prior

As discussed in Subsection 5.4, the choice of the prior may affect the performance of the Bayesian estimation. For instance, assuming that  $\Delta^{-1}$  has a Wishart distribution with scale matrix  $(\rho\Upsilon)$  and  $\rho$  degrees of freedom, Hsiao Pesaran and Tahmiscioglu (1999) note that the bias of both the empirical and hierarchical Bayes estimators of the regression coefficients can be sensitive to the specification of the prior scale matrix. Therefore, it is interesting to compare the performance of the hierarchical Bayes estimator with different prior choices. In a first specification, we use the same prior structure as in Hsiao, Pesaran and Tahmiscioglu (1999).<sup>22</sup> In a second specification, we set  $\Upsilon$  equal to the EM-REML estimator of the covariance matrix  $\Delta$  instead of the Swamy estimator.<sup>23</sup> Results are shown in Figure 4 and 5. By simply replacing the prior for  $\Upsilon$  with a more precise estimate of  $\Delta$ , obtained employing the EM-REML approach, the performances of the posterior mean of both the average effects

---

estimator of interest, given by the square root of  $(B-1)^{-1} \sum_{b=1}^B \left( \hat{\psi}_{k,(b)} - \bar{\hat{\psi}}_k \right)^2$ , where  $\bar{\hat{\psi}}_k = B^{-1} \sum_{b=1}^B \hat{\psi}_{k,(b)}$ , for  $k = 1, 2, 3$ .

<sup>21</sup>For the Mean Group estimation, the t-ratios are also appropriated. To facilitate comparison we only report the power functions computed using the Swamy type test, noting that in both cases the results are very similar.

<sup>22</sup>Under the assumption that the vector of average effects,  $\psi$ , has a prior distribution which is Normal with mean  $\mu$  and variance  $\Omega$ , the authors set  $\Omega^{-1} = 0$ ,  $\rho = 2$ , and choose  $\Upsilon$  equal to the Swamy estimate of  $\Delta$ .

<sup>23</sup>In both cases, we simulate a Markov chain of 6000 cycles, and discard the initial 1000 burn-in replications. Hsiao, Pesaran and Tahmiscioglu (1999) note that convergence is quickly achieved, and suggest using 3000 iterations.

and the variance components (in terms of bias and RMSE), notably improve, especially in small samples. This evidence confirms Kass and Wasserman (1996) argument that the prior choice can have heavy weight on the posterior when sample sizes are small.

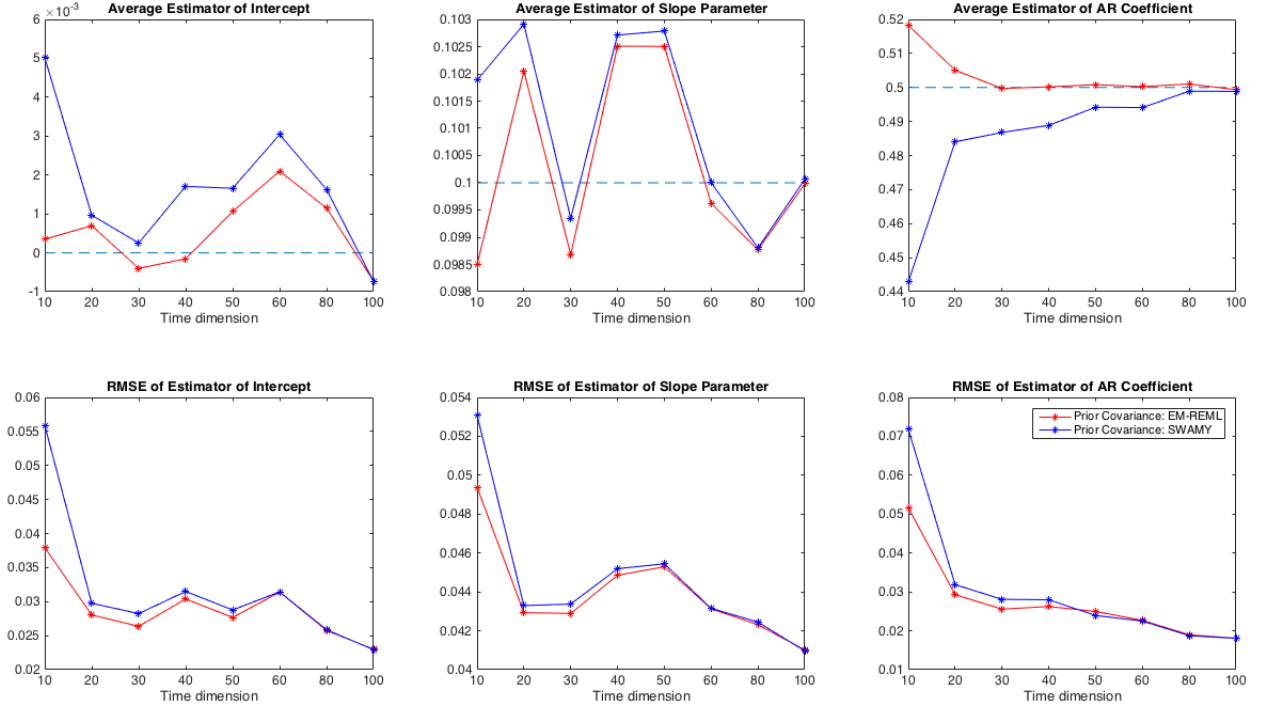


Figure 4: **Upper panel:** Posterior means for the intercept (left), slope (middle) and autoregressive parameter (right panel), averaged across 500 replications, are plotted for  $N = 30$  and  $T = \{10, 20, 30, 40, 50, 60, 80, 100\}$ . The dashed blue lines indicate the true values (used to simulate the data). Results in blue are obtained using priors as in Hsiao et al. (1999). Results in red are obtained using the EM-REML estimate of the random coefficient covariance as prior input. The distances between those lines and the one corresponding to the true value measure the bias of the estimators. **Lower panel:** the root mean square errors (RMSE) of the estimators are reported.

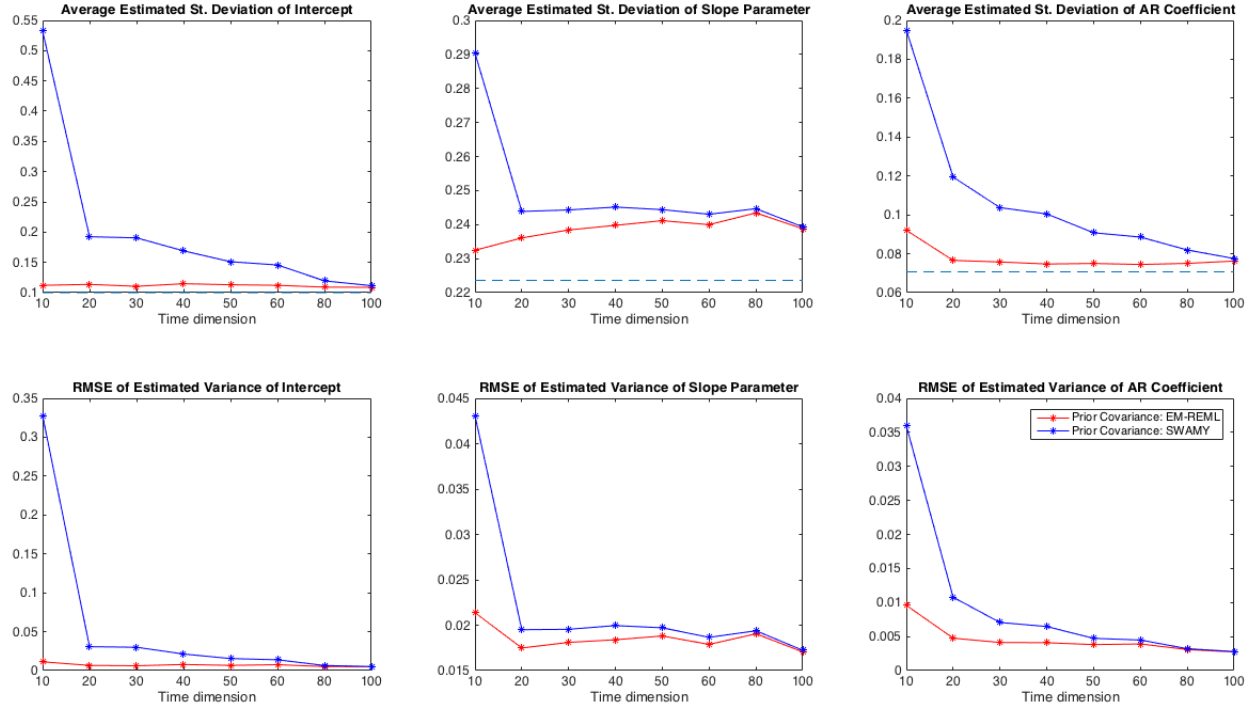


Figure 5: **Upper panel:** Posterior means for the variance of the intercept (left), slope (middle) and autoregressive parameter (right panel), averaged across 500 replications, are plotted for  $N = 30$  and  $T = \{10, 20, 30, 40, 50, 60, 80, 100\}$ . The dashed blue line indicates the true value (used to generate the data). Results in blue are obtained using priors as in Hsiao et al. (1999). Results in red are obtained using the EM-REML estimate of the random coefficient covariance. The distances between those lines and the one corresponding to the true value measure the bias of the estimators as prior input. **Lower panel:** the root mean square errors (RMSE) of the estimators are reported.

## 8 Application

Reinhart, Rogoff and Savastano (2003), studying sovereigns' credit histories since the early nineteenth century, argue that an important portion of middle-income countries has been "systematically" afflicted by what they call "debt intolerance". Even though their debt-to-GDP ratios are considerably lower than those of several high-income countries, these economies are considered to be riskier and unable to tolerate as much debt. We corroborate this argument by first showing that the response of sovereign spreads to changes in government debt (which we also refer to as the "sensitivity" of financial markets during episodes of debt growth) is highly heterogeneous. It is only statistically significant for a small subgroup of countries. We ask why this is so by modelling the sensitivity of spreads as function of

macroeconomic fundamentals and a set of explanatory variables which reflect the history of government debt and economic crises of various forms. We find that the more pervasive the phenomenon of serial default is (i.e. the weaker the reputation), the stronger the reaction of financial markets when debt increases. We quantify such reactions.

We depart from the literature on the determinants of sovereign spreads in several ways.<sup>24</sup> First, instead of considering only one group of countries (e.g. emerging markets), we collect quarterly data for a panel of 17 emerging market economies and 21 developed countries over 22 years (1994Q1-2015Q4).<sup>25</sup> Second, we consider a dynamic model. Third, given that we are comparing countries with very different characteristics, even within group, we allow for heterogeneity rather than pooling. The implications of neglected heterogeneity and dynamics can be severe. Pesaran and Smith (1995) show that if the DGP includes lagged values of the dependent variables among the explanatory variables, pooling give inconsistent and potentially highly misleading estimates of the coefficients when the latter differ across units. Haque, Pesaran and Sharma (2000) find that ignoring differences across countries can lead to overestimating the influence of certain factors. They argue that one can obtain highly significant, but spurious, nonlinear effects for some of the potential determinants, even though the country-specific regressions are linear.

Finally, the focus of this application is on understanding which factors determine the additional risk premium to charge during episodes of debt growth.

Assume that sovereign spreads are a function of debt-to-GDP ratio, a proxy for history of default and other macroeconomic fundamentals. Rather than looking at how spreads change with respect to one variable while debt-to-GDP and the remaining covariates are held constant (i.e. partial effect), we investigate which country characteristics significantly affect the magnitude of sovereign spreads' reaction to changes in debt. Studying the sensitivity of financial markets during episode of debt growth may help understand why emerging markets cannot borrow at level comparable to more developed economies without having to pay relatively high interest rates.

## 8.1 The Empirical Model

Following Edwards (1984), we assume that the spreads over U.S. (or Germany) Treasuries can be explained by a set of macroeconomic indicators. We focus on real GDP growth, inflation, and the growth rates of general gross government debt as a percentage of GDP. J.P. Morgan's Emerging Markets Bond Index Global (EMBI Global) is our measure of government bond

---

<sup>24</sup>See for instance, Akitoby and Stratmann (2008), Bellas et al. (2010), Edwards (1984), Eichengreen and Mody (2000) and Hilscher and Nosbusch (2010), among others.

<sup>25</sup>The panel is slightly unbalanced. The individual time observations vary between  $60 \leq T_i \leq 87$ . The choice of countries is dictated by the availability of data. The list of countries is reported in Appendix B.



yields for emerging markets.<sup>26</sup>

Because linear interdependencies may exist among these time series, we can assume they follow a VAR( $p$ ) process. Given that the spreads are observed at a daily frequency, it is reasonable to think that they react near-instantaneously to shocks and news. Therefore, considering the variables under study, we assume that the economy possesses a recursive structure where spreads are ordered last. The last equation of the recursive system can be written as

$$y_{it} = \phi_i y_{it-1} + x'_{it} \beta_{0i} + x'_{i,t-1} \beta_{1i} + \mu_i + \varepsilon_{it}, \quad (52)$$

for  $i = 1, \dots, N$  and  $t = 1, \dots, T$ ;  $y_{it}$  includes the first difference of sovereign spreads. The number of lags has been selected using the BIC criterion (averaged across units) since it results in more parsimonious model than the AIC. The panel data model in matrix notation can be written as in equation (2) where all the coefficients are random and follow (3). When doing parameter equality tests we set  $f_{1i} = 1$  for all  $i = 1, \dots, N$ , to then extend the analysis to the case where  $f_{1i}$  is a  $l \times 1$  vector of unit-specific explanatory variables.

## 8.2 Parameter Equality Tests

Before estimating the model, we employ some homogeneity tests to show that both the slope and the intercept parameters are heterogenous across countries. To test the null hypothesis  $H_0 : \psi_1 = \dots = \psi_N = \psi$  (i.e. to test whether the coefficient vectors  $\psi_i = (\mu_i, \beta'_{0i}, \phi_i, \beta'_{1i})'$  are constant across units), we can use the following test proposed by Swamy (1970):

$$F = \frac{1}{(N-1)} \sum_{i=1}^N F_i \sim F \left( K^*(N-1), \left( \sum_{i=1}^N T_i - NK^* \right) \right), \quad (53)$$

where

$$F_i = \frac{(\hat{\psi}_i - \hat{\psi})' Z_i' Z_i (\hat{\psi}_i - \hat{\psi})}{K^* \hat{\sigma}_{\varepsilon_i}^2},$$

and

$$\hat{\psi} = \left( \sum_{i=1}^N \frac{Z_i' Z_i}{\hat{\sigma}_i^2} \right)^{-1} \left( \sum_{i=1}^N \frac{Z_i' Z_i}{\hat{\sigma}_i^2} \hat{\psi}_i \right) = \left( \sum_{i=1}^N \frac{Z_i' Z_i}{\hat{\sigma}_i^2} \right)^{-1} \left( \sum_{i=1}^N \frac{Z_i' y_i}{\hat{\sigma}_i^2} \right).$$

$K^*$  is the dimension of  $\psi$ . The  $\hat{\psi}_i$ 's are obtained by estimating  $N$  time series separately by OLS. This test is appropriate in our case, since it should be used when  $T$  is large relative to  $N$ . For 296 and 2708 degrees of freedom, the F-value that leaves exactly 0.01 of the area under the F curve in the right tail of the distribution is smaller than 1.32.<sup>27</sup> Because our

<sup>26</sup>A description of the data is provided in Appendix B.

<sup>27</sup>The 1% significance level has been arbitrary chosen.

test has a value of 2.58, we are able to reject the null of homogenous slope and intercept parameters.

### 8.3 The Sensitivity of Spreads to Debt

We now explore why the sensitivity of spreads to debt differs significantly across countries by modelling the latter as a function of selected explanatory variables. We ask which factors influence financial markets decision when evaluating the credit worthiness of the borrower and setting interest rate during episodes of government debt growth.

Using Reinhart and Rogoff (2011) historical time series on countries creditworthiness and financial turmoil, we calculate the percentage of years (between 1980 and 2010) each country has been in default or restructuring on its domestic and external debt, the percentage of years with annual inflation of 20% or higher, and the percentage of years with annual depreciation vs US dollar of 15% or more. We then estimate equation (52) while allowing the coefficients to be a function of a common constant, and the percentage of years in default or restructuring domestic and external debt. Results are shown in Table 1.

Our results seem to suggest that history of repayment plays an important role: “bad” reputation leads to higher sensitivity of spreads to debt. A 1% increase in the percentage of years in default or restructuring domestic debt is associated with a 0.34% increase in the sensitivity of spread. As a consequence, relatively small increase in debt-to-GDP may lead to level of interest rates which can be difficult to tolerate. Although significant, the impact of our proxy for history of repayment of external debt is rather low, around 0.07 percent.

The above analysis is robust when augmenting the regression equation for the coefficients with additional explanatory variables. In particular, we include the percentage of years in which a country has faced an annual inflation rate of 20 percent or higher and the percentage of years in which an annual depreciation versus the US dollar (or another relevant anchor currency) of 15 percent or more occurs.<sup>28</sup> We also consider measures of macroeconomic fundamentals such as the average and standard deviation of real GDP growth, of rate of currency depreciation, of inflation and current account to GDP growth. Standard deviations over the sample period under considerations are used as measure of volatility. The standard deviation of the average growth rate of general gross government debt to GDP can be considered as a proxy for sudden increases in debt’s level.

In Table 2, we focus on the coefficients equation corresponding to the sensitivity of spreads to debt and report results from using different specifications. Including averages rather than volatility leads to very similar conclusions. Therefore, we do not report them. At least three conclusions can be drawn. First, a “good” reputation in financial markets matters. The percentage of years in defaults or restructuring on domestic debt have a statistically

---

<sup>28</sup>A detailed description of the data is provided by Reinhart and Rogoff (2009).

Table 1: Determinants of sensitivity of spreads: EM-REML Estimates.

	const.	% y-DomDef	% y-ExtDef
$c_i$	-0.017 (1.572)	0.596 (0.647)	-0.180 (0.897)
$\beta_0^{(gdp)}$	<b>-0.016*</b> (3.727)	-0.252 (0.364)	-0.082 (0.597)
$\beta_0^{(cpi)}$	0.008 (0.143)	0.624 (1.005)	-0.226 (1.502)
$\beta_0^{(debt)}$	-0.006 (1.311)	<b>0.344**</b> (5.998)	<b>0.068*</b> (3.264)
$\phi$	<b>0.112***</b> (8.035)	-0.326 (0.265)	-0.175 (0.501)
$\beta_1^{(gdp)}$	0.010 (1.060)	<b>-0.752*</b> (2.900)	<b>0.314***</b> (8.096)
$\beta_1^{(cpi)}$	<b>0.037*</b> (3.527)	-0.874 (2.201)	0.062 (0.128)
$\beta_1^{(debt)}$	0.003 (0.751)	-0.101 (0.616)	0.004 (0.014)

Swamy F-statistic (described in equation (46)) between parentheses. The critical values for a F distribution with 1 degree of freedom for the numerator, and  $N - 1$  for the denominator, associated with a significance level equal to 0.1, 0.05, and 0.01, are 2.84, 4.08, and 7.31 respectively. Symbols \*\*\*, \*\*, and \* denote significance (at least) at 1%, 5% and 10% respectively. Estimated standard errors are corrected for finite-sample bias, following Kenward and Roger (1997). “% y DomDef” (“% y ExtDef”) denotes the percentage of years in default or restructuring domestic (external) debt;  $\phi$  is the autoregressive coefficient;  $\beta^{(k)}$  is the sensitivity of spread to the  $k$ th variable.

and economically significant effect on the sensitivity of spreads across all the different specifications. Interestingly, domestic defaults have a larger impact than external ones. Our finding that domestic defaults play a significant role in explaining changes in the sensitivity of spreads is in line with Reinhart and Rogoff (2010) argument: “when ignored domestic debt obligations are taken into account, fiscal duress at the time of default is often revealed to be quite severe”. Second, country-specific macroeconomic indicators do not play any significant role in explaining the reactions of financial markets to an increase in debt. Contrary to the literature which emphasizes the role of volatility of macroeconomic aggregates in explaining sovereign credit risks, we do not find strong evidence that such variables affect markets when calculating the additional risk premium to charge in response to an increase in debt.<sup>29</sup> This

<sup>29</sup>Selected studies on the role of volatility in explaining sovereign defaults are: Eaton and Gersovitz (1981),

seems to suggest that markets decisions during episodes of debt growth may also be driven by sentiments (as defined by Eichengreen and Mody, 2000). At the same time, we have seen that a bigger reaction is usually associated with countries with a weak history of repayment.

Table 2: Determinants of sensitivity of spreads to government debt: EM-REML Estimates.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
<b>Constant</b>	-0.006 (1.311)	-0.008 (0.944)	-0.018 (1.719)	-0.016 (1.280)	0.001 (0.005)	-0.007 (0.181)	-0.019 (1.172)
<b>% y Dom Def</b>	<b>0.344</b> (5.998)	<b>0.328</b> (3.637)	<b>0.306</b> (4.592)	<b>0.308</b> (4.535)	<b>0.350</b> (6.085)	<b>0.340</b> (5.383)	
<b>% y Ext Def</b>	<b>0.068</b> (3.264)	0.026 (0.290)	0.058 (2.507)	<b>0.066</b> (2.990)	0.010 (0.074)	0.014 (0.122)	
<b>% y Curr Crisis</b>		-0.005 (0.006)					
<b>% y Infl Crisis</b>		0.066 (1.035)					
<b>Volatility FX</b>			0.003 (1.116)	0.003 (1.156)	0.001 (0.045)	0.003 (0.486)	0.006 (1.802)
<b>Volatility Debt/GDP</b>				-0.001 (0.313)	0.006 (2.652)	0.004 (1.023)	0.005 (1.889)
<b>Volatility Inflation</b>					0.011 (1.525)	0.005 (0.196)	0.008 (0.468)
<b>Volatility RGDP</b>					<b>-0.030</b> (5.798)	-0.016 (0.553)	-0.016 (0.532)
<b>Volatility CA/GDP</b>						-0.004 (0.392)	-0.008 (1.209)

Swamy F-statistic (described in equation (46)) between parentheses. The critical values for a F distribution with 1 degree of freedom for the numerator, and  $N-1$  for the denominator, associated with a significance level equal to 0.1, 0.05, and 0.01, are 2.84, 4.08, and 7.31 respectively. Estimated standard errors are corrected for finite-sample bias, following Kenward and Roger (1997). Bold values denotes statistical significance at 10% level or lower. “% y Curr Crisis” and “% y Infl Crisis” denote the percentage of years with annual inflation of 20% or higher and with an annual depreciation vs US dollar of 15% or more, respectively. “% y Dom Def” (“% y Ext Def”) denotes the percentage of years in default or restructuring domestic (external) debt.

To conclude, while it is common in the literature to find that certain macroeconomic fundamentals are significant predictors of sovereign spreads, we show that they are not significant determinants of the sensitivity of spreads to changes in sovereign debt. On the

Catao and Kapur (2006), and Hilscher and Nosbuch (2010).

contrary, reputation in financial markets is crucial.

## 9 Conclusion

This paper shows how to implement the EM algorithm to compute iteratively restricted maximum likelihood (REML) estimates of both fixed and random coefficients, as well as the variance components, in a wide class of heterogeneous panels. The proposed method has several benefits. First, the EM-REML approach yields an unbiased and more efficient estimator of the random coefficient covariance without running into the problem of negative definite matrices typically encountered in the Swamy type random coefficient models. This in turn leads to more accurate estimated standard errors and hypothesis tests. We also demonstrate that Lee and Griffiths (1979) approach to jointly estimate the random components and constant underlying parameters yield an estimator of the coefficients' covariance matrix which does not satisfy the law of total variance. This is not the case when employing the EM algorithm. Second, the latter allows us to make inference on the random effects' population. The EM approach should be considered as a valid alternative to Bayesian estimation in those cases in which the researcher wishes to make inference on the random effects' distribution while having little knowledge on what sensible priors might be. At the same time, it helps overcome one drawback of the Bayesian inference: when sample sizes are small (relative to the number of parameters being estimated), the prior choice will have a heavy weight on the posterior, which will consequently be far from being data dominated.

Monte Carlo experiments confirm that our approach performs relatively well in finite sample, in term of bias, root mean square errors, and power of tests.

Another contribution of this paper is to review in a coherent manner, some of the existing sampling and Bayesian methods commonly used to estimate random coefficient panel data models.

An empirical application is also presented. We investigate what causes the sensitivity of spreads to differ significantly across countries by modelling the latter as a function of macroeconomics fundamentals and a set of explanatory variables which reflect the history of government debt and economic crises of various forms. We ask which factors influence financial markets decision when evaluating the credit worthiness of the borrower and setting the risk premium during episodes of government debt growth. We find that while country-specific macroeconomic indicators (including underlying volatility) do not play any significant role in explaining the sensitivity of spreads to an increase in debt, history of repayment is crucial. "Bad" reputation leads to higher sensitivity of spreads to debt. An 1% increase in the percentage of years in default or restructuring domestic debt is associated with around 0.35% increase in the additional risk premium in response to an increase in debt. Our findings indicate that countries who have defaulted in the past may find it hard to finance government

expenditures by issuing new debt since relatively small increase in debt-to-GDP may lead to a raise in interest rates which may be difficult to tolerate. This helps explain why their debt-to-GDP ratios remain considerably lower than those of several high-income countries. The unanswered question is how to escape such a “trap”.

Table 3: Properties of EM-REML estimator as  $T$  gets large, for fixed  $N = 30$ 

$N = 30/T$	EM-REML							
	10	20	30	40	50	60	80	100
$Bias(\hat{c})$	<b>0.0015</b>	<b>0.0005</b>	<b>0.0001</b>	<b>-0.0002</b>	<b>0.0010</b>	<b>0.0020</b>	<b>0.0016</b>	<b>-0.0007</b>
$se\{Bias(\hat{c})\}$	0.0017	0.0012	0.0012	0.0013	0.0012	0.0014	0.0012	0.0011
$Bias(\hat{\beta})$	<b>-0.0033</b>	<b>0.0017</b>	<b>-0.0024</b>	<b>0.0021</b>	<b>0.0020</b>	<b>-0.0008</b>	<b>-0.0017</b>	<b>-0.0001</b>
$se\{Bias(\hat{\beta})\}$	0.0022	0.0019	0.0019	0.0020	0.0020	0.0019	0.0019	0.0018
$Bias(\hat{\phi})$	<b>0.0408</b>	<b>0.0104</b>	<b>0.0022</b>	<b>0.0037</b>	<b>0.0030</b>	<b>0.0022</b>	<b>0.0007</b>	<b>-0.0022</b>
$se\{Bias(\hat{\phi})\}$	0.0023	0.0013	0.0012	0.0012	0.0011	0.0010	0.0009	0.0009
$\ Bias(\hat{\psi})\ $	<b>0.0410</b>	<b>0.0106</b>	<b>0.0032</b>	<b>0.0042</b>	<b>0.0037</b>	<b>0.0030</b>	<b>0.0024</b>	<b>0.0023</b>
$RMSE(\hat{c})$	0.0391	0.0278	0.0269	0.0301	0.0278	0.0313	0.0263	0.0235
$RMSE(\hat{\beta})$	0.0497	0.0427	0.0435	0.0447	0.0450	0.0429	0.0422	0.0410
$RMSE(\hat{\phi})$	0.0659	0.0313	0.0259	0.0263	0.0252	0.0227	0.0196	0.0193
$Bias(v\hat{a}r(\gamma_1))$	<b>-0.0026</b>	<b>-0.0011</b>	<b>-0.0013</b>	<b>-0.0007</b>	<b>-0.0007</b>	<b>-0.0009</b>	<b>-0.0005</b>	<b>0.0003</b>
$se\{Bias(v\hat{a}r(\gamma_1))\}$	0.0005	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
$Bias(v\hat{a}r(\gamma_2))$	<b>-0.0122</b>	<b>-0.0046</b>	<b>-0.0030</b>	<b>-0.0025</b>	<b>-0.0016</b>	<b>-0.0020</b>	<b>-0.0002</b>	<b>-0.0021</b>
$se\{Bias(v\hat{a}r(\gamma_2))\}$	0.0008	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
$Bias(v\hat{a}r(\gamma_3))$	<b>-0.0002</b>	<b>-0.0011</b>	<b>-0.0010</b>	<b>-0.0010</b>	<b>-0.0008</b>	<b>-0.0008</b>	<b>-0.0004</b>	<b>0.0001</b>
$se\{Bias(v\hat{a}r(\gamma_3))\}$	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
$RMSE\{v\hat{a}r(\gamma_1)\}$	0.0118	0.0045	0.0046	0.0053	0.0047	0.0055	0.0040	0.0041
$RMSE\{v\hat{a}r(\gamma_2)\}$	0.0213	0.0146	0.0144	0.0143	0.0143	0.0138	0.0140	0.0133
$RMSE\{v\hat{a}r(\gamma_3)\}$	0.0062	0.0035	0.0032	0.0033	0.0031	0.0032	0.0026	0.0023
$Ratio\{se(\hat{c})\}$	1.01	0.98	0.96	1.01	0.98	1.00	0.98	1.01
$Ratio\{se(\hat{\beta})\}$	0.95	0.96	0.97	0.97	0.98	0.98	0.99	0.97
$Ratio\{se(\hat{\phi})\}$	1.56	1.12	1.08	1.15	1.06	1.11	1.01	1.03
$Accuracy\{se(\hat{c})\}$	0.87	0.94	0.93	0.95	0.93	0.90	0.89	0.97
$Accuracy\{se(\hat{\beta})\}$	0.89	0.97	0.95	0.93	0.92	0.96	0.98	0.98
$Accuracy\{se(\hat{\phi})\}$	1.06	0.95	0.95	1.04	0.88	1.03	0.95	0.94

The data generating process is described in equation (49) and (50), in Section 7. We assume that  $c_i \sim N(0, \sigma_c^2)$ ,  $\beta_i \sim N(0.1, \sigma_\beta^2)$ , and  $\phi_i \sim N(0.5, \sigma_\phi^2)$ , where  $(\sigma_c, \sigma_\beta, \sigma_\phi) = (0.1, 0.224, 0.07)$ . “se” stands for standard errors; RMSE indicates the root mean square errors. The Euclidean norm ( $\|\cdot\|$ ) is used as an overall measure of the bias.  $Ratio(se(\cdot))$  measures the ratio between the “infeasible” standard errors (which are obtained substituting the true values used to generate the DGP into equation (45)) and the estimated standard errors of the average effects.  $Accuracy(se(\cdot))$  denotes the accuracy of the estimated standard errors as approximations to the correct sampling standard deviation of the EM-REML estimator.

Table 4: Properties of Swamy estimator as  $T$  gets large, for fixed  $N = 30$ 

$N = 30/T$	Swamy							
	10	20	30	40	50	60	80	100
$Bias(\hat{c})$	<b>0.0211</b>	<b>0.0030</b>	<b>0.0018</b>	<b>0.0036</b>	<b>0.0018</b>	<b>0.0040</b>	<b>0.0025</b>	<b>-0.0007</b>
$se\{Bias(\hat{c})\}$	0.0044	0.0018	0.0018	0.0017	0.0015	0.0015	0.0012	0.0011
$Bias(\hat{\beta})$	<b>0.0025</b>	<b>0.0034</b>	<b>-0.0002</b>	<b>0.0032</b>	<b>0.0030</b>	<b>0.0002</b>	<b>-0.0013</b>	<b>-0.0001</b>
$se\{Bias(\hat{\beta})\}$	0.0026	0.0020	0.0020	0.0021	0.0021	0.0019	0.0019	0.0018
$Bias(\hat{\phi})$	<b>-0.1130</b>	<b>-0.0402</b>	<b>-0.0281</b>	<b>-0.0233</b>	<b>-0.0136</b>	<b>-0.0127</b>	<b>-0.0044</b>	<b>-0.0022</b>
$se\{Bias(\hat{\phi})\}$	0.0023	0.0014	0.0012	0.0012	0.0011	0.0010	0.0009	0.0009
$\ Bias(\hat{\psi})\ $	<b>0.1150</b>	<b>0.0405</b>	<b>0.0281</b>	<b>0.0237</b>	<b>0.0140</b>	<b>0.0133</b>	<b>0.0052</b>	<b>0.0023</b>
$RMSE(\hat{c})$	0.0998	0.0401	0.0403	0.0378	0.0343	0.0345	0.0278	0.0235
$RMSE(\hat{\beta})$	0.0581	0.0444	0.0446	0.0461	0.0459	0.0434	0.0424	0.0410
$RMSE(\hat{\phi})$	0.1240	0.0509	0.0395	0.0358	0.0282	0.0257	0.0203	0.0194
$Bias(v\hat{a}r(\gamma_1))$	<b>1.7745</b>	<b>0.1419</b>	<b>0.1501</b>	<b>0.0724</b>	<b>0.0530</b>	<b>0.0362</b>	<b>0.0133</b>	<b>0.0037</b>
$se\{Bias(v\hat{a}r(\gamma_1))\}$	0.0695	0.0042	0.0042	0.0020	0.0017	0.0011	0.0007	0.0004
$Bias(v\hat{a}r(\gamma_2))$	<b>0.1420</b>	<b>0.0209</b>	<b>0.0179</b>	<b>0.0126</b>	<b>0.0093</b>	<b>0.0081</b>	<b>0.0050</b>	<b>0.0013</b>
$se\{Bias(v\hat{a}r(\gamma_2))\}$	0.0042	0.0010	0.0010	0.0008	0.0008	0.0007	0.0007	0.0006
$Bias(v\hat{a}r(\gamma_3))$	<b>0.0825</b>	<b>0.0248</b>	<b>0.0152</b>	<b>0.0120</b>	<b>0.0075</b>	<b>0.0061</b>	<b>0.0033</b>	<b>0.0009</b>
$se\{Bias(v\hat{a}r(\gamma_3))\}$	0.0017	0.0006	0.0004	0.0004	0.0003	0.0003	0.0002	0.0001
$RMSE\{v\hat{a}r(\gamma_1)\}$	2.3571	0.1701	0.1769	0.0855	0.0648	0.0441	0.0199	0.0090
$RMSE\{v\hat{a}r(\gamma_2)\}$	0.1705	0.0307	0.0278	0.0218	0.0196	0.0180	0.0161	0.0143
$RMSE\{v\hat{a}r(\gamma_3)\}$	0.0904	0.0284	0.0178	0.0144	0.0099	0.0083	0.0052	0.0031
$Ratio\{se(\hat{c})\}$	7.34	2.90	3.06	2.13	1.96	1.65	1.33	1.08
$Ratio\{se(\hat{\beta})\}$	1.85	1.19	1.17	1.12	1.09	1.07	1.04	1.00
$Ratio\{se(\hat{\phi})\}$	2.00	1.61	1.46	1.33	1.25	1.20	1.14	1.02
$Accuracy\{se(\hat{c})\}$	2.53	1.93	1.98	1.60	1.51	1.35	1.13	1.03
$Accuracy\{se(\hat{\beta})\}$	1.49	1.16	1.11	1.04	1.00	1.05	1.02	1.01
$Accuracy\{se(\hat{\phi})\}$	1.35	1.30	1.19	1.15	1.06	1.13	1.06	0.93
% Negative Definite	0.81	0.71	0.72	0.72	0.60	0.59	0.47	0.16

The data generating process is described in equation (49) and (50), in Section 7. We assume that  $c_i \sim N(0, \sigma_c^2)$ ,  $\beta_i \sim N(0.1, \sigma_\beta^2)$ , and  $\phi_i \sim N(0.5, \sigma_\phi^2)$ , where  $(\sigma_c, \sigma_\beta, \sigma_\phi) = (0.1, 0.224, 0.07)$ . “se” stands for standard errors; RMSE indicates the root mean square errors. The Euclidean norm ( $\|\cdot\|$ ) is used as an overall measure of the bias.  $Ratio(se(\cdot))$  measures the ratio between the “infeasible” standard errors (which are obtained substituting the true values used to generate the DGP into equation (45)) and the estimated standard errors of the average effects. “% Negative Definite” measures the number of times (in percentage) the estimated random coefficients’ covariance matrix is negative definite.  $Accuracy(se(\cdot))$  denotes the accuracy of the estimated standard errors as approximations to the correct sampling standard deviation of the Swamy GLS estimator.



Table 5: Properties of Mean Group estimator as  $T$  gets large, for fixed  $N = 30$

$N = 30/T$	Mean Group							
	10	20	30	40	50	60	80	100
$Bias(\hat{c})$	<b>0.0626</b>	<b>0.0116</b>	<b>0.0103</b>	<b>0.0146</b>	<b>0.0091</b>	<b>0.0096</b>	<b>0.0064</b>	<b>-0.0001</b>
$se\{Bias(\hat{c})\}$	0.0112	0.0035	0.0036	0.0027	0.0025	0.0021	0.0017	0.0013
$Bias(\hat{\beta})$	<b>0.0033</b>	<b>0.0063</b>	<b>0.0019</b>	<b>0.0042</b>	<b>0.0041</b>	<b>0.0014</b>	<b>-0.0002</b>	<b>0.0018</b>
$se\{Bias(\hat{\beta})\}$	0.0036	0.0022	0.0023	0.0022	0.0022	0.0020	0.0020	0.0019
$Bias(\hat{\phi})$	<b>-0.2072</b>	<b>-0.0998</b>	<b>-0.0688</b>	<b>-0.0589</b>	<b>-0.0437</b>	<b>-0.0404</b>	<b>-0.0254</b>	<b>-0.0220</b>
$se\{Bias(\hat{\phi})\}$	0.0025	0.0015	0.0013	0.0013	0.0011	0.0010	0.0009	0.0008
$\ Bias(\hat{\psi})\ $	<b>0.2165</b>	<b>0.1006</b>	<b>0.0696</b>	<b>0.0608</b>	<b>0.0448</b>	<b>0.0415</b>	<b>0.0262</b>	<b>0.0220</b>
$RMSE(\hat{c})$	0.2575	0.0783	0.0820	0.0625	0.0569	0.0486	0.0387	0.0289
$RMSE(\hat{\beta})$	0.0803	0.0498	0.0509	0.0499	0.0494	0.0451	0.0436	0.0424
$RMSE(\hat{\phi})$	0.2146	0.1053	0.0744	0.0652	0.0501	0.0463	0.0325	0.0287
$Accuracy\{se(\hat{c})\}$	0.98	0.99	0.98	0.95	0.95	0.97	0.89	0.97
$Accuracy\{se(\hat{\beta})\}$	1.03	1.01	0.96	0.93	0.92	0.99	0.99	0.99
$Accuracy\{se(\hat{\phi})\}$	1.05	1.06	1.02	0.95	0.97	0.99	0.98	1.00

The data generating process is described in equation (49) and (50), in Section 7. We assume that  $c_i \sim N(0, \sigma_c^2)$ ,  $\beta_i \sim N(0.1, \sigma_\beta^2)$ , and  $\phi_i \sim N(0.5, \sigma_\phi^2)$ , where  $(\sigma_c, \sigma_\beta, \sigma_\phi) = (0.1, 0.224, 0.07)$ . “se” stands for standard errors; RMSE indicates the root mean square errors. The Euclidean norm ( $\|\cdot\|$ ) is used as an overall measure of the bias.  $Accuracy(se(\cdot))$  denotes the accuracy of the estimated standard errors as approximations to the correct sampling standard deviation of the Mean Group estimator.

# A Appendix

## A.1 Restricted Likelihood

### A.1.1 The Choice of $S_i$

**The projection matrix  $M_i$ .** One plausible choice for  $S_i$ , is the projection matrix:

$$M_i = I - W_i (W_i' W_i)^- W_i', \quad (54)$$

where  $(W_i' W_i)^-$  denotes the generalized inverse of  $W_i' W_i$ . The matrix  $M_i$  is of rank  $T - \underline{K}$ , with  $\underline{K} \leq \bar{K} < T$ , and satisfies  $M_i W_i = 0$ .  $M_i$  is symmetric and idempotent. As noted by Searle and Quaas (1978), its canonical form under orthogonal similarity is given by

$$U_i M_i U_i' = \begin{bmatrix} I_{T-\underline{K}} & 0 \\ 0 & 0 \end{bmatrix},$$

where  $U_i$  is an orthogonal matrix. Searle and Quaas (1978) defines  $A_i$  to be the first  $T - \underline{K}$  columns of  $U_i'$ . It follows that  $M_i = A_i A_i'$  and  $A_i' A_i = I$ . Premultiplying  $M_i$  by  $A_i$ , we get

$$M_i A_i = A_i, \quad A_i' M_i = A_i'. \quad (55)$$

Since  $U_i'$  is orthogonal and non-singular,  $A_i'$  has full rank and  $A_i' W_i = 0$ . Using (55), Searle and Quaas (1978) show that  $A_i (A_i' R_i A_i)^{-1} A_i'$  is the Moore-Penrose inverse of  $M_i R_i M_i$ :

$$(M_i R_i M_i)^+ = A_i (A_i' R_i A_i)^{-1} A_i'. \quad (56)$$

Given that  $A_i'$  has full row rank and  $R_i$  is positive definite, the inverse of  $A_i' R_i A_i$  exists.

**A generalization of  $M_i$ .** As shown in Searle and Quaas (1978), any linear combination of  $M_i$ ,  $S_i = J M_i$ , satisfies  $S_i W_i = 0$ . A generalization of  $M_i$  is

$$P_i = R_i^{-1} - R_i^{-1} W_i (W_i' R_i^{-1} W_i)^- W_i' R_i^{-1}, \quad (57)$$

satisfying  $P_i W_i = 0$ . From the definition of  $P_i$ , it follows that

$$\begin{aligned} R_i P_i &= I - W_i (W_i' R_i^{-1} W_i)^- W_i' R_i^{-1}, \\ P_i R_i &= I - R_i^{-1} W_i (W_i' R_i^{-1} W_i)^- W_i'. \end{aligned} \quad (58)$$

Therefore,

$$P_i R_i P_i = P_i, \quad (59)$$

and also  $(P_i R_i)^2 = P_i R_i$ . It follows that  $tr(P_i R_i) = rank(P_i R_i) = rank(P_i) = T - \underline{K}$ . Since  $P_i$  also satisfies  $P_i W_i = 0$ , we can choose  $S_i = P_i$ .

**Relationship between  $M_i$  and  $P_i$ .** Using (54) and the fact that  $P_i W_i = 0$ , it can be seen that

$$P_i M_i = P_i = M_i P_i. \quad (60)$$

Furthermore, post-multiplying (58) by  $M_i$  and using  $M_i W_i = 0$  and  $W_i' M_i = 0$ , we get  $P_i R_i M_i = M_i$ . Post-multiplying (60) by  $R_i M_i$

$$P_i M_i R_i M_i = P_i R_i M_i = M_i P_i R_i M_i = M_i^2 = M_i. \quad (61)$$

From (60) and (61), Searle and Quaas (1978) establish  $P_i$  as the Moore-Penrose inverse of  $M_i R_i M_i$ :

$$P_i = (M_i R_i M_i)^+. \quad (62)$$

Since  $(M_i R_i M_i)^+$  is unique, equations (56) and (62) imply that

$$P_i = (M_i R_i M_i)^+ = A_i (A_i' R_i A_i)^{-1} A_i'. \quad (63)$$

### A.1.2 Some Lemmas from Searle and Quaas (1978)

**Lemma 1.** Searle and Quaas (1978) show that  $S_i = F_i' A_i'$  for some non-singular  $F_i'$ . It follows that

$$\begin{aligned} S_i' (S_i R_i S_i')^{-1} S_i &= A_i F_i (F_i' A_i' R_i A_i F_i)^{-1} F_i' A_i' \\ &= A_i (A_i' R_i A_i)^{-1} A_i = P_i. \end{aligned} \quad (64)$$

where the last equality follows from (63).

**Lemma 2.** As shown in Lutkepohl (1996, pag. 50, eq. 6), if  $A$ ,  $B$ ,  $C$ , and  $D$  are  $(m \times m)$ ,  $(m \times n)$ ,  $(n \times m)$ , and  $(n \times n)$  matrices, respectively, then

$$\begin{aligned} \det \begin{bmatrix} A & B \\ C & D \end{bmatrix} &= |D| \cdot |A - BD^{-1}C| \quad \text{if } D \text{ nonsingular} \\ &= |A| \cdot |D - CA^{-1}B| \quad \text{if } A \text{ nonsingular} \end{aligned} \quad (65)$$

Using this property of the determinant, we can show that

$$|A_i R_i A_i'| = \frac{|R_i| \cdot |W_i' R_i^{-1} W_i|}{|W_i' W_i|}. \quad (66)$$

To prove the latter, let

$$\begin{bmatrix} A_i' \\ W_i' \end{bmatrix} R_i \begin{bmatrix} A_i & W_i \end{bmatrix} = \begin{bmatrix} A_i' R_i A_i & A_i' R_i W_i \\ W_i' R_i A_i & W_i' R_i W_i \end{bmatrix}.$$

Taking the determinant of both sides, we get

$$| R_i | \cdot \begin{vmatrix} A'_i A_i & A'_i W_i \\ W'_i A_i & W'_i W_i \end{vmatrix} = | A'_i R_i A_i | \cdot | W'_i R_i W_i - W'_i R_i A_i (A'_i R_i A_i)^{-1} A'_i R_i W_i |.$$

Using  $A'_i A_i = I$ , and  $A'_i W_i = 0$  and equation (63), we get

$$| R_i | | W'_i W_i | = | A'_i R_i A_i | \cdot | W'_i R_i W_i - W'_i R_i P R_i W_i |.$$

Substituting (58) into the latter equation and then using the following property of determinants,  $\det(AB) = \det(A) \cdot \det(B)$ , yields (66).

**Lemma 3.** Given that  $S_i = F'_i A_i$ , it can be shown that

$$| S_i R_i S'_i | = | F_i |^2 | A'_i R_i A_i |. \quad (67)$$

### A.1.3 Finding an expression for $L_{1i}$

Using (66) and (67), we have

$$\log | S_i R_i S'_i | = \mu + \log | R_i | + \log | W'_i R_i^{-1} W_i |, \quad (68)$$

where  $\mu$  includes the terms that do not involve the parameters of interest.

Furthermore, using (64), we get

$$\begin{aligned} (y_i - Z_i \gamma_i)' S'_i (S_i R_i S'_i)^{-1} S_i (y_i - Z_i \gamma_i) &= (y_i - Z_i \gamma_i)' P_i (y_i - Z_i \gamma_i) \\ &= \left( y_i - W_i \hat{\bar{\Gamma}} - Z_i \gamma_i \right)' R_i^{-1} \left( y_i - W_i \hat{\bar{\Gamma}} - Z_i \gamma_i \right). \end{aligned} \quad (69)$$

Substituting (68) and (69) into (14) yields (15).

**Proof of Equation (69).** Let  $\hat{\bar{\Gamma}}$  be the argument that minimizes  $\varepsilon'_i R_i^{-1} \varepsilon_i$ , where  $\varepsilon_i = y_i - W_i \bar{\Gamma} - \bar{Z}_i \gamma_i$  and  $R_i = \text{var}(\varepsilon_i)$ .<sup>30</sup> The solution to the problem is given by

$$\hat{\bar{\Gamma}} = \left( W'_i R_i^{-1} W_i \right)^{-1} W'_i R_i^{-1} \left( y_i - \bar{Z}_i \gamma_i \right).$$

---

<sup>30</sup>To make notation easier we focus on  $\varepsilon'_i R_i^{-1} \varepsilon_i$ . instead of  $\sum_{i=1}^N \varepsilon'_i R_i^{-1} \varepsilon_i$ . The same conclusion can be reached minimising the latter.

It follows that

$$\begin{aligned} y_i - W_i \hat{\Gamma} - \bar{Z}_i \gamma_i &= y_i - W_i (W_i' R_i^{-1} W_i)^{-} W_i' R_i^{-1} (y_i - \bar{Z}_i \gamma_i) - \bar{Z}_i \gamma_i \\ &= R_i P_i y_i - R_i P_i \bar{Z}_i \gamma_i. \end{aligned}$$

Therefore, using (59) and after a few computations, we get

$$\begin{aligned} \left( y_i - W_i \hat{\Gamma} - \bar{Z}_i \gamma_i \right)' R_i^{-1} \left( y_i - W_i \hat{\Gamma} - \bar{Z}_i \gamma_i \right) &= \left( y_i' P_i R_i - \gamma_i' \bar{Z}_i' P_i R_i \right) R_i^{-1} \left( R_i P_i y_i - R_i P_i \bar{Z}_i \gamma_i \right) \\ &= y_i' P_i y_i - y_i' P_i \bar{Z}_i \gamma_i - \gamma_i' \bar{Z}_i' P_i y_i + \gamma_i' \bar{Z}_i' P_i \bar{Z}_i \gamma_i \\ &= \left( y_i - \bar{Z}_i \gamma_i \right)' P_i \left( y_i - \bar{Z}_i \gamma_i \right). \end{aligned}$$

#### A.1.4 Finding an expression for $L_{2i}$ .

**The Choice of  $Q_i$ .** It can be shown that  $Q_i = W_i' R_i^{-1}$  satisfies  $\text{cov}(S_i y_i, Q_i y_i) = 0$ , and therefore is a plausible choice to obtain  $L_{2i}$ . We first compute the covariance conditional on  $\gamma_i$ , to then show that the unconditional covariance is equal to zero.

$$\begin{aligned} \text{cov}(S_i y_i, Q_i y_i \mid \gamma_i) &= E(S_i y_i y_i' Q_i' \mid \gamma_i) - E(S_i y_i \mid \gamma_i) E(y_i' Q_i' \mid \gamma_i) \\ &= S_i E(y_i y_i' \mid \gamma_i) Q_i' - (S_i \bar{Z}_i \gamma_i) (\bar{\Gamma}' W_i' + \gamma_i' \bar{Z}_i') R_i^{-1} W_i, \end{aligned} \quad (70)$$

where  $E(S_i y_i \mid \gamma_i) = S_i \bar{Z}_i \gamma_i$ , since  $S_i W_i = 0$ .

Substituting

$$S_i E(y_i y_i' \mid \gamma_i) Q_i' = S_i \text{var}(\varepsilon_i) Q_i' = S_i R_i R_i^{-1} W_i = S_i W_i = 0,$$

and

$$\begin{aligned} (S_i \bar{Z}_i \gamma_i) (\bar{\Gamma}' W_i' + \gamma_i' \bar{Z}_i') R_i^{-1} W_i &= S_i \bar{Z}_i \gamma_i \bar{\Gamma}' W_i' R_i^{-1} W_i \\ &\quad + S_i \bar{Z}_i \gamma_i \gamma_i' \bar{Z}_i' R_i^{-1} W_i \end{aligned}$$

into (70), we get

$$\begin{aligned} \text{cov}(S_i y_i, Q_i y_i \mid \gamma_i) &= -S_i \bar{Z}_i \gamma_i \bar{\Gamma}' W_i' R_i^{-1} W_i \\ &\quad - S_i \bar{Z}_i \gamma_i \gamma_i' \bar{Z}_i' R_i^{-1} W_i. \end{aligned} \quad (71)$$

Using the Law of Total Covariance, the unconditional covariance can be obtained from

$$\begin{aligned} \text{cov}(S_i y_i, Q_i y_i) &= E[\text{cov}(S_i y_i, Q_i y_i \mid \gamma_i)] \\ &\quad + \text{cov}(E(S_i y_i \mid \gamma_i), E(Q_i y_i \mid \gamma_i)). \end{aligned} \quad (72)$$

Taking expectation of both sides of (71), we get

$$E[\text{cov}(S_i y_i, Q_i y_i \mid \gamma_i)] = -S_i \bar{Z}_i \Delta \bar{Z}_i' R_i^{-1} W_i, \quad (73)$$

since  $\gamma_i \sim N(0, \Delta)$ . Moreover,

$$\begin{aligned} \text{cov}(E(S_i y_i | \gamma_i), E(Q_i y_i | \gamma_i)) &= E \left[ S_i \bar{Z}_i \gamma_i (W_i' R_i^{-1} W_i \bar{\Gamma} + W_i' R_i^{-1} \bar{Z}_i \gamma_i)' \right] \\ &\quad - E[E(S_i y_i | \gamma_i)] E[E(Q_i y_i | \gamma_i)] \\ &= S_i \bar{Z}_i \Delta \bar{Z}_i' R_i^{-1} W_i. \end{aligned} \quad (74)$$

Therefore, substituting (73) and (74) into (72) we can show that  $\text{cov}(S_i y_i, Q_i y_i) = 0$ .

## A.2 Best Linear Unbiased Prediction

**Conditional Mean and Variance.** Under the assumption that  $y_i$  and  $\gamma_i$  are jointly normally distributed, the conditional expectation of  $\gamma_i$  given the data is

$$\begin{aligned} \hat{\gamma}_i = E(\gamma_i | y_i) &= E(\gamma_i) + \text{cov}(\gamma_i, y_i) [\text{var}(y_i)]^{-1} [y_i - E(y_i)] \\ &= \kappa' V_i^{-1} (y_i - W_i \bar{\Gamma}), \end{aligned} \quad (75)$$

where  $E(\gamma_i) = 0$ , by assumption,  $E(y_i) = W_i \bar{\Gamma}$ ,  $V_i = \text{var}(y_i) = \bar{Z}_i \Delta \bar{Z}_i' + R_i$ , and  $\kappa' = \text{cov}(\gamma_i, y_i) = \Delta \bar{Z}_i'$ . The conditional variance of  $\gamma_i$  is

$$\begin{aligned} \text{var}(\gamma_i | y_i) &= \text{var}(\gamma_i) - \text{cov}(\gamma_i, y_i) [\text{var}(y_i)]^{-1} \cdot \text{cov}(y_i, \gamma_i) \\ &= \text{var}(\gamma_i) - \kappa' V_i^{-1} \kappa. \end{aligned} \quad (76)$$

As suggested in Pawitan (2001), using a simple matrix identity we can write

$$\begin{aligned} \Delta \bar{Z}_i' [\bar{Z}_i \Delta \bar{Z}_i' + R_i]^{-1} &= \left\{ \left( \bar{Z}_i' R_i^{-1} \bar{Z}_i + \Delta^{-1} \right)^{-1} \left( \bar{Z}_i' R_i^{-1} \bar{Z}_i + \Delta^{-1} \right) \right\} \\ &\quad \cdot \Delta \bar{Z}_i' [\bar{Z}_i \Delta \bar{Z}_i' + R_i]^{-1} \\ &= \left( \bar{Z}_i' R_i^{-1} \bar{Z}_i + \Delta^{-1} \right)^{-1} \cdot \bar{Z}_i' R_i^{-1}. \end{aligned} \quad (77)$$

This result is used to derive the second equality in equation (20) and to obtain equation (21).

**Properties.** Following Henderson (1984, Chap. 5), it can be shown that:

(i)  $\hat{\gamma}_i$  is an unbiased predictor of  $\gamma_i$ :

$$\begin{aligned} E(\hat{\gamma}_i) &= E \left[ \kappa' V_y^{-1} (y_i - W_i \bar{\Gamma}) \right] \\ &= \kappa' V_y^{-1} [E(y_i) - W_i \bar{\Gamma}] = E(\gamma_i), \end{aligned} \quad (78)$$

since  $E(y_i) = W_i \bar{\Gamma}$ .

- (ii)  $cov(\hat{\gamma}_i, \gamma_i) = var(\hat{\gamma}_i)$ , from which it follows that  $var(\hat{\gamma}_i - \gamma_i) = var(\gamma_i) - var(\hat{\gamma}_i)$ .  
(iii) the BLUP maximizes the correlation between  $\hat{\gamma}_i$  and  $\gamma_i$ .  
Finally, note that

$$\begin{aligned}
var(\hat{\gamma}_i) &= var\left[\kappa'V_i^{-1}(y_i - W_i\bar{\Gamma})\right] = \kappa'V_i^{-1}\kappa \\
&= \Delta_i\bar{Z}_i'\left(\bar{Z}_i\Delta\bar{Z}_i' + R_i\right)^{-1}\bar{Z}_i\Delta \\
&= \left(\bar{Z}_iR_i^{-1}\bar{Z}_i + \Delta^{-1}\right)^{-1}\bar{Z}_iR_i^{-1}\bar{Z}_i\Delta
\end{aligned} \tag{79}$$

### A.3 Expectation Step

**E-step for  $L_{2i}$ .** As suggested in Pawitan (2001), we can write

$$E_{\theta^{(b-1)}}(\varepsilon_i' H_i \varepsilon_i \mid y_i) = Tr[H_i E_{\theta^{(b-1)}}(\varepsilon_i \varepsilon_i' \mid y_i)]. \tag{80}$$

To find  $E_{\theta^{(b-1)}}(\varepsilon_i \varepsilon_i' \mid y_i)$ , recall that for a random vector  $x$ , with mean  $\mu_x$  and variance  $V_x$ ,  $var(x) = E(xx') - E(x)E(x')$ , from which it follows  $E(xx') = V_x + \mu_x\mu_x'$ . Therefore,

$$E_{\theta^{(b-1)}}(\varepsilon_i \varepsilon_i' \mid y_i) = V_{\varepsilon_i} + \hat{\varepsilon}_i \hat{\varepsilon}_i', \tag{81}$$

where

$$\begin{aligned}
\hat{\varepsilon}_i &= E_{\theta^{(b-1)}}(\varepsilon_i \mid y_i) = E_{\theta^{(b-1)}}(y_i - W_i\bar{\Gamma} - \bar{Z}_i\gamma_i \mid y_i) \\
&= y_i - W_i\bar{\Gamma} - \bar{Z}_i\hat{\gamma}_i^{(b)},
\end{aligned}$$

and

$$\begin{aligned}
V_{\varepsilon_i} &= var(\varepsilon_i \mid y_i; \theta^{(b-1)}) = var(y_i - W_i\bar{\Gamma} - \bar{Z}_i\gamma_i \mid y_i, \theta^{(b-1)}) \\
&= \bar{Z}_i V_{\gamma_i}^{(b)} \bar{Z}_i',
\end{aligned} \tag{82}$$

with  $\hat{\gamma}_i^{(b)} = E_{\theta^{(b-1)}}(\gamma_i \mid y_i)$  and  $V_{\gamma_i}^{(b)} = var(\gamma_i \mid y_i, \theta^{(b-1)})$ .

Substituting (81) into (80) yields

$$\begin{aligned}
E_{\theta^{(b-1)}}(\varepsilon_i' H_i \varepsilon_i \mid y_i) &= Tr\left(H_i Z_i V_{\gamma_i}^{(b)} Z_i'\right) + Tr(H_i \hat{\varepsilon}_i \hat{\varepsilon}_i') \\
&= Tr\left(Z_i' H_i Z_i V_{\gamma_i}^{(b)}\right) + \hat{\varepsilon}_i' H_i \hat{\varepsilon}_i.
\end{aligned}$$

We can now write

$$\begin{aligned}
Q_{2i} = E_{\theta^{(b-1)}}(L_{2i} \mid y_i) &= c_4 - \frac{1}{2} \log |W_i' R_i^{-1} W_i| \\
&\quad - \frac{1}{2} Tr\left(Z_i' H_i Z_i V_{\gamma_i}^{(b)}\right) - \frac{1}{2} \hat{\varepsilon}_i' H_i \hat{\varepsilon}_i.
\end{aligned}$$

Using a similar expedient, we can obtain  $Q_{1i}$  and  $Q_{3i}$ .

## A.4 Estimation of $\Delta$

An estimator of  $\Delta$  can be obtained by maximizing  $\sum_{i=1}^N Q_{3i}$ , where  $Q_{3i}$  is defined in (24), with respect to  $\Delta$ . Before proceeding, we report a few results of matrices differentiation shown in Lutkepohl (1996).

1.  $X$  ( $m \times m$ ) nonsingular,  $a, b$  ( $m \times 1$ ):

$$\frac{\partial a' X^{-1} b}{\partial X} = -(X^{-1})' a b' (X^{-1})'. \quad (83)$$

2.  $X$  ( $m \times m$ ) nonsingular,  $A, B$  ( $m \times m$ ):

$$\frac{\partial \text{tr}(AX^{-1}B)}{\partial X} = -(X^{-1}BAX^{-1})'. \quad (84)$$

3.  $X$  ( $m \times m$ ),  $\det(X) > 0$ :

$$\frac{\partial \ln |X|}{\partial X} = (X')^{-1}. \quad (85)$$

Therefore,

$$\frac{\partial Q_{3i}}{\partial \Delta} = \underbrace{-\Delta^{-1}}_{(85)} + \underbrace{\Delta^{-1} V_{\gamma_i}^{(b)} \Delta^{-1}}_{(84)} + \underbrace{\Delta^{-1} \hat{\gamma}_i^{(b)} \hat{\gamma}_i^{(b)' \Delta^{-1}}}_{(83)} = 0, \quad (85)$$

The solution to  $(\partial \sum_{i=1}^N Q_{3i} / \partial \Delta) = 0$  is given by

$$\hat{\Delta} = \frac{1}{N} \sum_{i=1}^N \{V_{\gamma_i} + \hat{\gamma}_i \hat{\gamma}_i'\}. \quad (86)$$

**Unbiased Estimator.** It can be shown that

$$\hat{\Delta} = \frac{1}{N} \sum_{i=1}^N \{V_{\gamma_i} + \hat{\gamma}_i \hat{\gamma}_i'\} \quad (87)$$

is an unbiased estimator of  $\Delta$  since

$$\begin{aligned} E(\hat{\Delta}) &= N^{-1} \sum_{i=1}^N \{E(\hat{\gamma}_i \hat{\gamma}_i') + E(V_{\gamma_i})\} \\ &= N^{-1} \sum_{i=1}^N \left\{ E \left[ \kappa' V_i^{-1} (y_i - W_i \bar{\Gamma}) (y_i - W_i \bar{\Gamma})' V_i^{-1} \kappa \right] + \Delta - \kappa' V_i^{-1} \kappa \right\} \\ &= N^{-1} \sum_{i=1}^N \left\{ \kappa' V_i^{-1} \kappa - \kappa' V_i^{-1} \kappa \right\} + \Delta = \Delta. \end{aligned}$$



## A.5 Hypothesis Testing

### A.5.1 Covariance of Estimator of Fixed Coefficients

Noting that  $V = \text{var}(y)$  has the linear form

$$V = \sum_{s=1}^{\bar{r}} \vartheta_s \Pi_s,$$

the adjusted estimator of the small sample variance-covariance matrix of  $\hat{\Gamma}$  is given by

$$\hat{\Phi}_A = \hat{\Phi} + 2\hat{\Lambda}, \quad (88)$$

where

$$\begin{aligned} \hat{\Lambda} &= \hat{\Phi} \left\{ \sum_{s=1}^r \sum_{j=1}^r \Upsilon_{sj} \left( \Xi_{sj} - \Sigma_s \hat{\Phi} \Sigma_j \right) \right\} \hat{\Phi}, \\ \Sigma_s &= - \sum_{i=1}^N W_i' V_i^{-1} \Pi_{s,i} V_i^{-1} W_i, \\ \Xi_{sj} &= \sum_{i=1}^N W_i' V_i^{-1} \Pi_{s,i} V_i^{-1} \Pi_{j,i} V_i^{-1} W_i. \end{aligned}$$

An approximation to  $\Upsilon$ , the variance-covariance matrix of  $\hat{\vartheta}$ , can be obtained from the inverse of the expected information matrix  $I_E$ , where

$$2 \{I_E\}_{sj} = \text{tr} \left( \sum_{i=1}^N V_i^{-1} \Pi_{s,i} V_i^{-1} \Pi_{j,i} \right) - 2 \text{tr} (\Phi \Xi_{sj}) + \text{tr} (\Phi \Sigma_s \Phi \Sigma_j).$$

Detailed derivations are provided by Alnosaier (2007).

### A.5.2 Assessing the Errors of Estimation for the Unit-Specific Coefficients

The variance-covariance matrix of the predictor of (5), is given by

$$\begin{aligned} \text{var}(\hat{\psi}_{1i} - \psi_{1i}) &= F_{1i} \text{var}(\hat{\Gamma}) F_{1i}' + \text{var}(\hat{\gamma}_i - \gamma_i) + F_{1i} \text{cov}(\hat{\Gamma} - \bar{\Gamma}, \hat{\gamma}_i - \gamma_i) \\ &\quad + \left[ F_{1i} \text{cov}(\hat{\Gamma} - \bar{\Gamma}, \hat{\gamma}_i - \gamma_i) \right]', \end{aligned} \quad (89)$$

where

$$\begin{aligned} \text{cov}(\hat{\Gamma} - \bar{\Gamma}, \hat{\gamma}_i - \gamma_i) &= \text{cov}(\hat{\Gamma} - \bar{\Gamma}, \hat{\gamma}_i) - \text{cov}(\hat{\Gamma} - \bar{\Gamma}, \gamma_i) \\ &= -\Phi W_i' V_i^{-1} \bar{Z}_i \Delta, \end{aligned}$$

since  $\text{cov}(\hat{\Gamma}, \hat{\gamma}_i) = 0$ , and  $\text{cov}(y_i, \hat{\Gamma}) = \text{var}(\hat{\Gamma}) W_i'$ .

## B Data

### B.1 List of Countries

**Advanced Economies:** Australia (AU), Austria (OE), Belgium (BG), Canada (CN), Denmark (DK), Finland (FN), France (FR), Greece (GR), Iceland (IC), Ireland (IR), Italy (IT), Japan (JP), Netherlands (NL), New Zealand (NZ), Norway (NW), Portugal (PT), Singapore (SP), Spain (ES), Sweden (SD), Taiwan (TW), United Kingdom (UK).

**Emerging Market and Developing Economies:** Argentina (AG), Brazil (BR), Chile (CL), China (CH), Croatia (CT), Hungary (HN), India (IN), Malaysia (MY), Mexico (MX), Peru (PE), Philippines (PH), Poland (PO), Russia (RS), South Africa (SA), Thailand (TH), Turkey (TK), Venezuela (VE).

The classification of countries follows from IMF, World Economic Outlook, October 2015 (pag.187-188).

### B.2 Data Sources

**Bond Yields:** J.P. Morgan EMBI Global, OECD Main Economic Indicators, Eurostat (for DK, GR, LX, and PT), and national authorities (OE, IN, IT, SP, SD, TW).

**Bond Spreads:** for all European countries but Iceland, the bond spread is measured against German long-term government bond yields. For the remaining countries, the bond spread is measured against US long-term government bond yields.

**Current Accounts:** OECD Main Economic Indicators, Oxford Economics, and national Central Banks (LV, PE).

**Government Debt:** Oxford Economics, Eurostat (LV, LX, SJ), and Bank for International Settlements (IR, IS, NZ, PE).

**CPI inflation:** IMF - International Financial Statistics, OECD Main Economic Indicators (AG, CL, CH, SX), and Oxford Economics (SP, TW, TH).

**Real GDP:** Oxford Economics, national authorities (IS, LV, LX, NZ, PE), and OECD Main Economic Indicators (SJ).

**Exchange Rates:** IMF - International Financial Statistics, OECD Main Economic Indicators, and Oxford Economics (TW).

**Financial History:** Historical time series on countries creditworthiness and financial turmoil are obtained from Reinhart and Rogoff (2009, 2011).

## References

- [1] B. Akitoby and T. Stratmann. Fiscal policy and financial markets. *The Economic Journal*, 118(533):1971–1985, 2008.
- [2] W. S. Alnosaier. *Kenward-Roger approximate F test for fixed effects in mixed linear models*. PhD thesis, 2007.
- [3] T. Amemiya and W. A. Fuller. A comparative study of alternative estimators in a distributed lag model. *Econometrica, Journal of the Econometric Society*, pages 509–529, 1967.
- [4] T. W. Anderson and C. Hsiao. Estimation of dynamic models with error components. *Journal of the American statistical Association*, 76(375):598–606, 1981.
- [5] T. W. Anderson and C. Hsiao. Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18(1):47–82, 1982.
- [6] B. Baltagi. *Econometric analysis of panel data*. John Wiley & Sons, 2008.
- [7] B. Baltagi, G. Bresson, and A. Pirotte. To pool or not to pool. In Matyas and Sevestre, editors, *The Econometrics of Panel data*, volume 3rd edition, pages 517–546. Springer-Verlag, Berlin, 2008.
- [8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B*, pages 1–38, 1977.
- [9] J. Eaton and M. Gersovitz. Debt with potential repudiation: Theoretical and empirical analysis. *The Review of Economic Studies*, 48(2):289–309, 1981.
- [10] S. Edwards. LDC’s foreign borrowing and default risk: An empirical investigation. *The American Economic Review*, 74(4):726–734, 1984.
- [11] B. Eichengreen and A. Mody. What explains changing spreads on emerging-market debt: Fundamentals or market sentiment? *NBER Working Paper Series*, page 6408, 1998.
- [12] R. J. Friedrich. In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science*, pages 797–833, 1982.
- [13] A.-M. Fuertes and R. Smith. *Panel Time Series*. CEMMAP, 2016.
- [14] J. D. Hamilton. *Time series analysis*. Princeton University Press, 1994.

- [15] U. N. Haque, M. H. Pesaran, and S. Sharma. Neglected heterogeneity and dynamics in cross-country savings regressions. In J. Krishnakumar and E. Rouchetti, editors, *Panel Data Econometrics - Future Directions: Papers in Honour of Prof. Balestra*, pages 53–82. Elsevier Science, 2000.
- [16] C. R. Harvey and Y. Liu. Rethinking performance evaluation. *National Bureau of Economic Research*, (No. w22134), 2016.
- [17] D. A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.
- [18] F. Hayashi. *Econometrics*. Princeton University Press, 2000.
- [19] C. R. Henderson. *Applications of Linear Models in Animal Breeding*, volume 1. 1983.
- [20] J. Hilscher and Y. Nosbusch. Determinants of sovereign risk: Macroeconomic fundamentals and the pricing of sovereign debt. *Review of Finance*, pages 235–262, 2010.
- [21] J. P. Hobert and G. Casella. The effect of improper priors on gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91(436):1461–1473, 1996.
- [22] C. Hsiao. *Analysis of panel data*. Cambridge University Press, 2003.
- [23] C. Hsiao, T. W. Appelbe, and C. R. Dineen. A general framework for panel data models with an application to canadian customer-dialed long distance telephone service. *Journal of Econometrics*, 59(1-2):63–86, 1993.
- [24] C. Hsiao and M. H. Pesaran. Random coefficient panel data models. In Matyas and Sevestre, editors, *The Econometrics of Panel Data*, 2008.
- [25] C. Hsiao, M. H. Pesaran, and A. K. Tahmiscioglu. Bayes estimation of short-run coefficients in dynamic panel data models. In C. Hsiao, Lahiri, L. K., and M. Pesaran, editors, *Analysis of Panels and Limited Dependent Variable Models*, 1999.
- [26] R. N. Kacker and D. A. Harville. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79(388):853–862, 1984.
- [27] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

- [28] M. G. Kenward and J. H. Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, pages 983–997, 1997.
- [29] G. Koop, D. J. Poirier, and J. L. Tobias. *Bayesian econometric methods*. Cambridge University Press, 2007.
- [30] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.
- [31] L.-F. Lee and W. E. Griffiths. *The prior likelihood and best linear unbiased prediction in stochastic coefficient linear models*. Citeseer, 1979.
- [32] D. V. Lindley and A. F. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B*, pages 1–41, 1972.
- [33] H. Lutkepohl. *Handbook of matrices*. John Wiley & Sons, 1997.
- [34] G. Maddala. Generalized least squares with an estimated variance covariance matrix. *Econometrica: Journal of the Econometric Society*, pages 23–33, 1971.
- [35] G. S. Maddala, R. P. Trost, H. Li, and F. Joutz. Estimation of short-run and long-run elasticities of energy demand from panel data using shrinkage estimators. *Journal of Business & Economic Statistics*, 15(1):90–100, 1997.
- [36] P. Mazodier and A. Trognon. Heteroscedasticity and stratification in error components models. In *Annales de l’INSEE*, pages 451–482. JSTOR, 1978.
- [37] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.
- [38] A. Mian and A. Sufi. *House of debt: How they (and you) caused the Great Recession, and how we can prevent it from happening again*. University of Chicago Press, 2015.
- [39] Y. Mundlak. On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, pages 69–85, 1978.
- [40] A. Pagan. Two stage and related estimators and their applications. *The Review of Economic Studies*, 53(4):517–538, 1986.
- [41] H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

- [42] Y. Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.
- [43] H. Pesaran, R. Smith, and K. S. Im. Dynamic linear models for heterogenous panels. In *The econometrics of panel data*, pages 145–195. Springer, 1996.
- [44] M. H. Pesaran and R. Smith. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 68(1):79–113, 1995.
- [45] I. Petrova, M. M. G. Papaioannou, and M. D. Bellas. *Determinants of emerging market sovereign bond spreads: fundamentals vs financial stress*. Number 10-281. International Monetary Fund, 2010.
- [46] W. C. Randolph. A transformation for heteroscedastic error components regression models. *Economics Letters*, 27(4):349–354, 1988.
- [47] C. M. Reinhart. This time is different chartbook: country histories on debt, default, and financial crises. Technical report, National Bureau of Economic Research, 2010.
- [48] C. M. Reinhart and K. S. Rogoff. *This time is different: eight centuries of financial folly*. Princeton University Press, 2009.
- [49] C. M. Reinhart and K. S. Rogoff. From financial crash to debt crisis. *The American Economic Review*, 101(5):1676–1706, 2011.
- [50] K. Reinhart, Carmen Rogoff, , and M. Savastano. Debt intolerance. *Brookings Papers on Economic Activity*, 1:1–74, 2003.
- [51] N. Roy. Is adaptive estimation useful for panel models with heteroskedasticity in the individual specific error component? some monte carlo evidence. *Econometric Reviews*, 21(2):189–203, 2002.
- [52] S. Searle, R. Quaas, et al. A notebook on variance components: A detailed description of recent methods of estimating variance components, with applications in animal breeding. 1978.
- [53] A. F. Smith. A general bayesian linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 67–75, 1973.
- [54] P. A. Swamy. Efficient inference in a random coefficient regression model. *Econometrica: Journal of the Econometric Society*, pages 311–323, 1970.

- [55] P. A. V. B. Swamy. *Statistical inference in random coefficient regression models*, volume 55. Springer Science & Business Media, 2012.
- [56] L. Trapani and G. Urga. Optimal forecasting with heterogeneous panels: A monte carlo study. *International Journal of Forecasting*, 25(3):567–586, 2009.